Ciencias cognitivas



David J. Chalmers

# La mente consciente

*En busca de una teoría fundamental*





gedisa
editorial

# LA MENTE CONSCIENTE

*En busca de una teoría fundamental*

*por*

David J. Chalmers

gedisa
editorial

Título del original en inglés: *The Conscious Mind. In Search of a Fundamental Theory*
Publicado por Oxford University Press


Esta traducción de *The Conscious Mind* editado originalmente en inglés en 1966 se publica por acuerdo con Oxford University Press Inc.

Traducción de José A. Álvarez



Primera edición: septiembre de 1999. Barcelona


Muntaner, 460, entlo., 1ª
Tel. 93-201 60 00
08006 Barcelona. España
correo-e: gedisa@gedisa.com
http://www.gedisa.com

ISBN: 84-7432-692-3
Depósito legal: B. 36.386-1999

Impreso por Limpergraf
Mogoda, 29-31. 08210 Barberà del Vallès (Barcelona)

Impreso en España
*Printed in Spain*

# Agradecimientos

Me interesé por primera vez en la conciencia y en el problema mente-cuerpo cuando era estudiante de matemática en la Universidad de Adelaida. Las conversaciones con algunas personas, especialmente Paul Barter, Jon Baxter, Ben Hambly y Paul McCann, ayudaron a darles forma a mis ideas. Ya entonces, el tema me parecía un problema fascinante como ninguno. Se me hacía poco razonable que alguien pudiese tener una ocupación de tiempo completo para pensar en algo que resultaba tan divertido.

Más tarde, como estudiante graduado en Oxford, descubrí que, en lugar de la matemática, la mente ocupaba siempre mis pensamientos, por lo cual decidí cambiar de campo y, finalmente, de continente. Muchas personas fueron pacientes y me apoyaron durante esta época difícil, en especial Michael Atiyah, Michael Dummet y Robin Fletcher. Agradezco también a todos aquellos que estuvieron sometidos a escuchar la que fuera en ese momento mi más reciente teoría de la conciencia; las ideas de este libro son descendientes lejanos de aquellas.

Mi decisión de mudarme a la Universidad de Indiana para obtener una formación en filosofía, ciencia cognitiva e inteligencia artificial fue una de las mejores que tomé. Debo agradecer especialmente a Doug Hofstadter; fue su obra la que me introdujo por primera vez en los misterios de la mente cuando yo era joven, y fue el ambiente estimulante y confortable de su laboratorio de investigación, el Center for Research on Concepts and Cognition, el que permitió que esas ideas se desarrollasen. Aunque él discrepa con muchas de las ideas de este libro, me gustaría pensar que en algún nivel mi trabajo es fiel al espíritu intelectual de su obra.

Escribí la primera versión de este libro (entonces conocido como *Toward a Theory of Consciousness*) en un impetuoso período de seis

meses entre 1992 y 1993. En esa época tuve provechosas discusiones con algunas personas de Indiana: todos del CRCC, especialmente Bob French y Liane Gabora, y muchos de otros departamentos, incluyendo a Mike Dunn, Rob Goldstone, Anil Gupta, Jim Hettmer, Jerry Seligman y Tim van Gelder. Gracias también a los miembros del grupo de discusión sobre la conciencia, en la trastienda de Nick's, por muchas agradables conversaciones los lunes a la tarde.

Una beca de investigación de dos años de McDonnell para estudiar filosofía, neurociencia y psicología en la Universidad de Washington me proporcionó otro ambiente estimulante, además de la oportunidad de experimentar la paradoja de Zenón en la terminación de este libro. Agradezco a la James S. McDonnell Foundation por su apoyo, a todos los participantes en mi seminario de grado sobre la conciencia por las discusiones que me ayudaron a pulir el libro, y a varias personas por las conversaciones y comentarios, incluyendo a Morten Christiansen, Andy Clark, Jason Clevenger, Peggy DesAutels, Pepa Toribio, y Tad Zawidzki.

En los últimos dos años mantuve una enorme cantidad de provechosas conversaciones y correspondencia sobre el material de este libro. Entre muchos otros, debo agradecer a Jon Baxter, Ned Block, Alex Byrne, Francis Crick, Dan Dennett, Eric Dietrich, Avi Elitzur, Matthew Elton, Owen Flanagan, Stan Franklin, Liane Gabora, Güven Güzeldere, Chris Hill, Terry Horgan, Steve Horst, Frank Jackson, Jaegwon Kim, Christof Koch, Martin Leckey, Dave Leising, Kerry Levenberg, Joe Levine, David Lewis, Barry Loewer, Bill Lycan, Paul McCann, Daryl McCullough, Brian McLaughlin, Thomas Metzinger, Robert Miller, Andrew Milne, John O'Leary-Hawthorne, Joseph O'Rourke, Calvin Ostrum, Rhett Savage, Aaron Sloman, Leopold Stubenberg y Red Watson. Agradezco también a muchos otros que me sería difícil nombrar por interesantes conversaciones sobre la conciencia en general. Una nota especial de agradecimiento a Norton Nelkin, quien me devolvió su copia del manuscrito llena de útiles comentarios no mucho antes de que muriese de linfoma. Lo extrañaremos.

Mis deudas filosóficas más amplias son muchas. Mis primeros puntos de vista sobre la conciencia los desarrollé, en lo fundamental, por mí mismo, pero se enriquecieron mucho gracias a las lecturas sobre el tema. Rápidamente uno descubre que cualquier idea tal vez haya sido expresada por otra persona. Entre los pensadores recientes, Thomas Nagel, Frank Jackson y Joseph Levine hicieron mucho por destacar las perplejidades de la conciencia; el trabajo de estos autores abarca gran parte de los mismos temas que mis primeros capítulos. También, mi trabajo se superpone en algunos puntos con el de Ned

Block, Robert Kirk y Michael Lockwood. El marco de referencia metafísico que desarrollo en el capítulo 2 tiene una gran deuda con el trabajo de Terry Horgan, Saul Kripke y David Lewis, entre otros; Frank Jackson desarrolló en forma independiente una teoría similar, que presentó en sus excelentes conferencias John Locke de 1995. Las ideas de Daniel Dennett, Colin McGinn, John Searle y Sydney Shoemaker representaron en todo momento desafíos estimulantes.

Mis mayores deudas las tengo con Gregg Rosenberg, por memorables conversaciones y valiosos intercambios de ideas; con Lisa Thomas, por un libro sobre zombis y su apoyo moral; con Sharon Wahl, por su experta corrección y cálida amistad, y, fundamentalmente, con mis tres padres por su apoyo y estímulo. Y gracias también a todos mis qualia, y al ambiente responsable de producirlos, por su constante inspiración.

Cuando estaba concluyendo este libro, recibí en un restaurante una galleta de la fortuna que decía: “Su vida estará colmada de encantadores misterios”. Hasta ahora así fue y por ello estoy agradecido.

# Índice

# INTRODUCCIÓN

## Tomar en serio a la conciencia

La conciencia es un gran misterio. Es, tal vez, el mayor obstáculo pendiente en nuestra búsqueda de una comprensión científica del universo. La ciencia física aún es incompleta, pero tenemos de ella una buena comprensión; la ciencia de la biología se deshizo de muchos antiguos misterios acerca de la naturaleza de la vida. Existen resquicios en nuestra comprensión de estos campos, pero no parecen imposibles de abordar. Tenemos una buena idea de la forma que podrían tener las soluciones de estos problemas; solamente necesitamos precisar los detalles.

Incluso en la ciencia de la mente se hicieron muchos progresos. Trabajos recientes en la ciencia cognitiva y en la neurociencia nos llevan a una mejor comprensión de la mente humana y de los procesos que la rigen. Seguramente, no tenemos muchas teorías detalladas de la cognición, pero los detalles no pueden estar muy lejos de nuestro alcance. La conciencia, sin embargo, sigue siendo tan desconcertante como siempre. Todavía nos resulta algo totalmente misterioso que la causalidad de la conducta esté acompañada de una vida interior subjetiva.

Tenemos buenas razones para pensar que la conciencia surge en sistemas físicos como el cerebro, pero tenemos poca idea de cómo es que surge o por qué existe. ¿Cómo podría un sistema físico como el cerebro ser también un *experimentador*? ¿Por qué debiera haber *algo que es como* un sistema de esta clase? Las teorías científicas actuales casi no abordan las preguntas realmente difíciles sobre la conciencia. No sólo carecemos de una teoría detallada; estamos totalmente a oscuras acerca de cómo encaja la conciencia en el orden natural.

En los últimos años aparecieron muchos libros y artículos sobre la conciencia, por lo que podría pensarse que estamos haciendo algún progreso. Sin embargo, en un examen más atento, puede verse que la

mayor parte de estos trabajos dejan sin tratar los problemas más difíciles acerca de la conciencia. Frecuentemente, esos trabajos encaran lo que podría llamarse los problemas "fáciles" de la conciencia: ¿cómo procesa el cerebro los estímulos ambientales? ¿cómo integra la información? ¿cómo producimos informes sobre nuestros estados internos? Estas son preguntas importantes, pero su respuesta no significa resolver el problema difícil: ¿Por qué todo este procesamiento está acompañado por una vida interna que experimentamos? A veces se ignora por completo esta pregunta; a veces se la posterga para algún momento futuro y a veces simplemente se la declara resuelta. Pero, en cada caso, nos queda la sensación de que el problema central sigue siendo tan enigmático como siempre.

Esta perplejidad no debe ser motivo de desesperanza; más bien, hace que el problema de la conciencia sea uno de los desafíos intelectuales más excitantes de nuestro tiempo. Debido a que la conciencia es, al mismo tiempo, tan fundamental y tan incomprendida, una solución al problema podría afectar profundamente nuestra concepción del universo y de nosotros mismos.

Yo soy optimista acerca de la conciencia: creo que con el tiempo lograremos una teoría de ella, y este libro es un intento de encontrarla. Pero la conciencia no es un problema común; para lograr algún progreso, lo primero que debemos hacer es enfrentar las cuestiones que hacen que el problema sea tan difícil. Luego podremos avanzar hacia una teoría, sin pestañear y con una buena idea de la tarea que nos aguarda.

En este libro, no resuelvo el problema de la conciencia de una vez por todas, pero intento encauzarlo. Trato de aclarar cuáles son los problemas, sostengo que los métodos estándar de la neurociencia y la ciencia cognitiva no sirven para enfrentarlos y luego intento avanzar.

En el desarrollo de mi concepción de la conciencia, he tratado de obedecer a un número de restricciones. La primera y la más importante es *tomar en serio a la conciencia*. El modo más fácil de desarrollar una "teoría" de la conciencia es negar su existencia, o redefinir el fenómeno que requiere explicación como algo que no la necesita. Esto, por lo general, lleva a una teoría elegante, pero el problema no desaparece. En este libro supongo que la conciencia existe, y que es inaceptable redefinir el problema diciendo que sólo se trata de explicar en qué forma se realizan ciertas funciones cognitivas o conductuales. Esto es lo que quiero decir por tomar en serio a la conciencia.

Algunos dicen que la conciencia es una "ilusión", pero yo no tengo idea de lo que esto significa. Me parece que estamos más seguros de

la existencia de la experiencia consciente de lo que estamos de cualquier otra cosa en el mundo. En ocasiones intenté intensamente convencerme a mí mismo de que en realidad no hay nada allí, que la experiencia consciente es vacua, una ilusión. Hay algo de atractivo en esta noción que tantos filósofos, en todas las épocas explotaron, pero al final resulta totalmente insatisfactoria. Cuando me encuentro absorto en una sensación del color naranja, *algo ocurre*. Hay algo que requiere explicación, aun luego de haber aclarado los procesos de discriminación y acción: se trata de la *experiencia*.

Es verdad, yo no puedo *probar* que haya un problema ulterior, precisamente porque no puedo probar que la conciencia existe. Lo que sabemos sobre la conciencia es más inmediato de lo que sabemos de cualquier otra cosa, de modo que una "demostración" es inapropiada. Lo más que puedo hacer es proporcionar argumentos cada vez que ello sea posible y refutar los argumentos del otro bando. No puedo negar que, en algún punto, esto requiere apelar a la intuición; pero todos los argumentos necesitan en algo de la intuición, y yo he tratado de ser lo más claro posible acerca de las intuiciones involucradas en los míos.

Podría considerarse que esta es una gran línea divisoria en el estudio de la conciencia. Si usted sostiene que una respuesta a los problemas "fáciles" explica todo lo que debe ser explicado, entonces tiene un tipo de teoría; si usted sostiene que existe un problema "difícil" ulterior, entonces tiene otro tipo. Pasado un cierto punto, es difícil *argumentar* a través de esta línea divisoria, y las discusiones frecuentemente se reducen a golpear sobre la mesa. A mí me parece obvio que hay algo más que necesita una explicación; a otros, les parece aceptable que no lo haya. (Encuestas informales sugieren que la proporción es de dos o tres a uno en favor del primer enfoque, con una tasa bastante constante entre los investigadores y estudiantes de una variedad de campos.) Quizá, simplemente debamos aprender a vivir con esta división básica.

Este libro puede ser de interés intelectual para aquellos que piensan que en realidad no hay ningún problema, pero está dirigido principalmente a aquellos que sienten el problema como propio. En la actualidad ya tenemos una idea bastante buena del tipo de teoría que obtenemos cuando suponemos que no hay un problema ulterior. En este trabajo, he intentado explorar qué es lo que se puede inferir si se supone que sí existe un problema. La verdadera tesis del libro es que, *si* se toma en serio a la conciencia, se debería llegar a la posición que formulo.

La segunda restricción que adopté es *tomar en serio a la ciencia*. No intenté cuestionar las teorías científicas actuales en dominios sobre los que estas tienen autoridad. Al mismo tiempo, no temí

arriesgarme en áreas en las que las opiniones de los científicos están tan carentes de fundamentos como las de cualquiera. Por ejemplo, no niego que el mundo físico esté causalmente cerrado o que la conducta pueda explicarse en términos físicos; pero si algún físico o un científico cognitivo sugiere que la conciencia puede explicarse en términos físicos, esto es meramente una expresión de deseos que no se basa en la teoría actual, y la cuestión sigue abierta. De este modo, traté de que mis ideas se mantuviesen *compatibles* con la ciencia contemporánea, pero no las limité a lo que los científicos contemporáneos encuentran aceptable.

La tercera restricción consiste en aceptar que la conciencia es un fenómeno natural sometido al dominio de las leyes naturales. Si esto es así, entonces debería haber *alguna* teoría científica correcta de la conciencia, podamos o no formularla. Parece difícil poner en duda que la conciencia sea un fenómeno natural: es una parte extraordinariamente sobresaliente de la naturaleza que surge en la especie humana y muy probablemente en muchas otras especies. Y tenemos todas las razones para creer que los fenómenos naturales están sujetos a leyes naturales fundamentales; sería muy extraño que la conciencia no lo estuviese. Esto no significa que las leyes naturales acerca de la conciencia sean como las leyes en otros dominios, o incluso que sean *leyes físicas*. Podrían ser de una clase muy distinta.

El problema de la conciencia está instalado de un modo inestable en la frontera entre la ciencia y la filosofía. Yo diría que es propiamente un tema científico: es un fenómeno natural como el movimiento, la vida y la cognición, y reclama una explicación lo mismo que ellos. Pero no está abierto a la investigación mediante los métodos científicos usuales. La metodología científica ordinaria tiene dificultades para captarlo, y una causa importante de esto son las dificultades para observar el fenómeno. Fuera del caso de primera persona, es difícil encontrar datos. Esto no significa que ningún dato externo pueda ser relevante, pero primero debemos llegar a una comprensión filosófica coherente antes de poder justificar la relevancia de los datos. De esta forma, el problema de la conciencia podría ser un problema científico que requiere métodos filosóficos de comprensión antes de que podamos despegar.

En este libro, llego a conclusiones que algunas personas podrían considerar "anticientíficas": argumento que una explicación reductiva de la conciencia es imposible, y sostengo, incluso, una forma de dualismo. Pero esto es sólo parte del proceso científico. Ciertos tipos de explicaciones no resultan apropiadas, de modo que en su lugar necesitamos aceptar otros. Todo lo que aquí digo es compatible con los resultados de la ciencia contemporánea; nuestra imagen

del mundo natural se amplía, no se subvierte. Y esta ampliación permite una teoría naturalista de la conciencia que sería imposible sin ella. Me parece que *ignorar* los problemas de la conciencia sería anticientífico; es coherente con el espíritu científico enfrentarlos directamente. A aquellos que sospechan que la ciencia requiere del materialismo, les pido que esperen y vean.

Debo hacer notar que las conclusiones de este trabajo son *conclusiones* en el sentido más fuerte de la palabra. Por temperamento, me inclino poderosamente por la explicación reductiva materialista, y no poseo ninguna fuerte inclinación espiritual o religiosa. Durante algunos años conservé la esperanza de lograr una teoría materialista; abandoné esa esperanza con bastante renuencia. Finalmente me resultó evidente que estas conclusiones eran obligatorias para cualquiera que quisiera tomar en serio a la conciencia. El materialismo es una cosmovisión hermosa y atractiva, pero para poder ofrecer una concepción de la conciencia debemos ir más allá de los recursos que este provee.

En este momento, sin embargo, estoy casi feliz con las conclusiones. No parecen tener ninguna consecuencia temible y permiten, en cambio, un modo de pensamiento y teorización sobre la conciencia que parece más satisfactorio en casi todos los aspectos. Y la expansión de la cosmovisión científica ha tenido un efecto positivo, al menos en mí: hizo que el universo parezca un lugar más interesante.

Este libro tiene cuatro partes. En la primera, planteo los problemas y defino un marco general dentro del cual estos pueden encararse. El capítulo 1 es una introducción al tema de la conciencia en el que analizamos algunos de los diferentes conceptos vinculados, extraemos un sentido en el que la conciencia es realmente interesante, y realizamos una descripción preliminar de la sutil relación que mantiene con el resto de la mente. El capítulo 2 desarrolla un marco metafísico y explicativo dentro del cual se plasma gran parte del resto del análisis. ¿Qué significa que un fenómeno sea explicado reductivamente, o que sea físico? Este capítulo expone una concepción de estas cuestiones, y se concentra en la noción de superveniencia. Sostengo que hay buenas razones para creer que *casi* todo en el mundo puede explicarse reductivamente; pero la conciencia podría ser una excepción.

Habiendo superado estos preliminares, la segunda parte se concentra en la irreducibilidad de la conciencia. En el capítulo 3 sostengo que los métodos estándar de la explicación reductiva no pueden dar cuenta de la conciencia. También hago una crítica a las diversas concepciones reductivas formuladas por los investigadores

de la neurociencia, la ciencia cognitiva y otras disciplinas. Esta no es sólo una conclusión negativa: se deduce también que una teoría satisfactoria de la conciencia debería ser de una nueva clase, *no reductiva*, de teoría. En el capítulo 4 llevamos las cosas un paso más allá argumentando que el materialismo es falso y que una forma de dualismo es verdadera, y esbozamos la forma general que podría adoptar una teoría no reductiva de la conciencia. El capítulo 5 es fundamentalmente defensivo: considera algunos problemas evidentes de mi teoría que involucran la relación entre la conciencia y nuestros *juicios* sobre la misma, y sostiene que no plantean dificultades fatales.

En la tercera parte, paso a considerar una teoría positiva de la conciencia. Cada uno de los tres capítulos de esta parte desarrolla un componente de una teoría positiva. El capítulo 6 se concentra en la "coherencia" entre la conciencia y los procesos cognitivos, y formula un cierto número de vínculos sistemáticos entre los dos. Utilizo estos vínculos para analizar y fundamentar el papel central de la neurociencia y la ciencia cognitiva en la explicación de la conciencia humana. En el capítulo 7 se analiza la relación entre la conciencia y la organización funcional y se utilizan experimentos mentales para argumentar que la conciencia es un "invariante organizativo": esto es, que todo sistema con la organización funcional correcta tendrá la misma clase de experiencia consciente, sin que importe de qué está hecho. En el capítulo 8 considero cuál sería la forma de una *teoría fundamental* de la conciencia, y sugiero que podría involucrar una relación estrecha entre la conciencia y la información. Este es, de lejos, el capítulo más especulativo, pero es probable que en este punto se necesite algo de especulación para poder progresar.

Los dos últimos capítulos son el postre. En ellos aplico lo anterior a cuestiones centrales en la fundamentación de la inteligencia artificial y la mecánica cuántica. En el capítulo 9 defiendo la tesis de una "inteligencia artificial fuerte": la implementación de un programa computacional apropiado dará origen a una mente consciente. En el capítulo 10 considero la sorprendente cuestión de cómo debería interpretarse la mecánica cuántica y utilizo las ideas sobre la conciencia desarrolladas en los capítulos previos para dar apoyo a una interpretación "no colapsante" de la teoría.

Es posible que el material negativo sea el que produzca las mayores reacciones, pero mi verdadero objetivo es positivo: yo quiero obtener una *teoría* de la conciencia que funcione. Cuando ingresé en la filosofía, me sorprendió descubrir que la mayor parte del debate sobre la conciencia se concentraba en la cuestión de si había un problema o no, o en la de si era un fenómeno físico o no, y que la

cuestión de construir teorías parecía haberse dejado de lado. Las únicas "teorías" parecían haber sido formuladas por aquellos que (a mi entender) no tomaban en serio a la conciencia. En la actualidad, disfruto de las complejidades del debate ontológico tanto como cualquiera, pero mi principal objetivo sigue siendo una teoría detallada. Si algunas de las ideas en este libro les resultan útiles a otros para construir una mejor teoría, el intento habrá valido la pena.

Esta obra pretende ser un trabajo serio de filosofía, pero me propuse que fuese accesible también a los no filósofos. Dentro de mi audiencia hipotética siempre estuvo mi propio yo estudiante de hace diez años: espero haber escrito un libro que él hubiera apreciado. Existen algunas secciones que son filosóficamente técnicas. Estas están marcadas con un asterisco (*), y los lectores pueden sentirse en libertad de saltearlas. El material más técnico se encuentra en los capítulos 2 y 4. El apartado 4 del primero, los apartados 2 y 3 del segundo y el apartado final del capítulo 5 contienen intrincadas cuestiones de semántica filosófica. Podría resultarle útil leer, aunque sólo fuera superficialmente, otras secciones con asterisco para formarse una idea de lo que allí expongo. Por lo general, coloqué el material o los comentarios especialmente técnicos sobre la literatura filosófica en notas al pie. El único concepto técnico crucial para el libro es el de superveniencia, que se introduce en el comienzo del capítulo 2. Este concepto tiene un nombre intimidante, pero expresa una idea muy natural, y una buena comprensión de él le ayudará a situar en el lugar justo a las cuestiones centrales. Gran parte del material posterior de este capítulo puede saltearse en una primera lectura, aunque usted podría querer volver a él más tarde para aclarar cuestiones a medida que estas surgen.

Para una breve excursión que evite los tecnicismos, lea el capítulo 1, lea superficialmente las primeras partes del capítulo 2 como conocimiento general, luego lea todo el capítulo 3 (pasando rápidamente por el apartado 1, si es necesario) para recoger los principales argumentos en contra de la explicación reductiva, y, en el primer y último apartado del capítulo 4, las consideraciones centrales sobre el dualismo. La lectura del comienzo del capítulo 6 es útil para entender la forma básica del enfoque positivo. Del material positivo, el capítulo 7 es, quizás, el más independiente y también el más divertido, con experimentos mentales fáciles de comprender que involucran cerebros de silicio; aquellos que disfrutan de las especulaciones osadas e imprecisas pueden encontrar interesante el capítulo 8. Finalmente, los capítulos 9 y 10 pueden ser interesantes para cualquiera que se preocupe por las cuestiones involucradas.

Un par de anotaciones filosóficas es oportuno. La literatura filosófica sobre la conciencia es bastante asistemática: está conformada por líneas aparentemente independientes que tratan cuestiones relacionadas sin hacer contacto entre sí. Intenté imponer alguna estructura al desorden, mediante un marco unificador en el cual las diversas cuestiones metafísicas y explicativas se aclaran. Gran parte de la discusión en la literatura puede trasladarse a este marco general sin pérdida, y espero que esta estructura haga patente las profundas relaciones entre una serie de diferentes cuestiones.

Este trabajo quizá sea inusual en el hecho de evitar lo más posible la noción filosófica de identidad (entre estados mentales y físicos, por ejemplo) en favor de la noción de superveniencia. Encuentro que, por lo general, las discusiones formuladas en términos de identidad arrojan más confusión que luz sobre las cuestiones claves y, con frecuencia, permiten eludir las dificultades centrales. La superveniencia, en cambio, parece proporcionar un marco ideal dentro del cual pueden encararse las cuestiones cruciales. Para evitar una filosofía floja, sin embargo, debemos concentrarnos en la *fortaleza* de la conexión de superveniencia: ¿está respaldada por la necesidad lógica, por la necesidad natural o por algo más? Existe un amplio consenso acerca de que la conciencia, en cierto sentido, superviene a lo físico; el verdadero problema es cuán estrecha es la conexión. Las discusiones que ignoran estas cuestiones modales suelen eludir las preguntas más difíciles acerca de la conciencia. Las personas que sean escépticas acerca de esas nociones modales también lo serán respecto de todo mi análisis, pero creo que no hay ningún otro modo satisfactorio de enmarcar las cuestiones.

Para mí, uno de los placeres de trabajar en este libro provino del modo como el problema de la conciencia se extendió para hacer contacto con profundos problemas de muchas otras áreas de la ciencia y la filosofía. Pero el alcance y profundidad del problema también lo hacen humillante. Soy muy consciente de que en casi cualquier punto de este libro habría más para decir, y que en muchos sitios apenas arañé la superficie. Sin embargo, espero, aunque más no sea en forma mínima, haber sugerido que es posible hacer progresos en el problema de la conciencia sin negar su existencia o reducirla a algo que no es. El problema es fascinante y el futuro, excitante.

# PARTE I

# FUNDAMENTOS

*No*. Xia se detuvo, girando en espiral hacia él con movimientos lentos. Sus ojos de menta helada se ensancharon. *Te encuentras en peligro aquí.* El pánico hizo palidecer su rostro mientras miraba hacia la casa. *Ve a casa ahora. Antes de que sea demasiado tarde. Y encuentra el antídoto.*

*¿Qué* tipo *de antídoto?*

Xia desapareció detrás de los enebros, mientras su mensaje final estallaba en la mente de Joey como la detonación de un petardo: *El antídoto para el veneno zombi.*

Dian Curtis Regan, *My Zombie Valentine*

# 1

# Dos conceptos de la mente

## 1. ¿Qué es la conciencia?

La experiencia consciente es, al mismo tiempo, lo más familiar del mundo y lo más misterioso. De ninguna otra cosa tenemos un conocimiento más directo que de la conciencia, pero no es claro en absoluto cómo reconciliarla con todo el resto de lo que sabemos. ¿Por qué existe? ¿Qué hace? ¿Cómo puede surgir a partir de la grumosa materia gris? Conocemos a la conciencia de una forma mucho más íntima de lo que conocemos al resto del mundo, pero comprendemos a este último mucho mejor de lo que comprendemos a la conciencia.

La conciencia puede ser sorprendentemente intensa. Es el más vívido de los fenómenos; nada es más real para nosotros. Pero puede ser frustrantemente diáfana: al hablar acerca de la experiencia consciente es muy difícil definir el tema. El *The International Dictionary of Psychology* ni siquiera intenta caracterizarla directamente:

> *Conciencia*: Tener percepciones, pensamientos y sentimientos; percatación. El término es imposible de definir excepto en términos que son ininteligibles sin una captación de lo que la conciencia significa. Muchos caen en la trampa de confundir la conciencia con la autoconciencia; para ser consciente sólo es necesario percatarse del mundo externo. La conciencia es un fenómeno fascinante pero huidizo: es imposible especificar qué es, qué hace o por qué evolucionó. Aún no se ha escrito nada que valga la pena leerse. (Sutherland, 1989)

Casi cualquiera que haya pensado intensamente sobre la conciencia sentirá una cierta simpatía hacia estos sentimientos. La conciencia es tan intangible que aun este intento limitado por definir-

la podría ser cuestionable: es posible argumentar que existen percepciones y pensamientos que no son conscientes, como lo atestiguan las nociones de percepción subliminal y de pensamiento inconsciente. Lo que es central para la conciencia, al menos en el sentido más interesante, es la *experiencia*. Pero esto no es una definición. A lo sumo es una aclaración.

Intentar definir la experiencia consciente en términos de nociones más primitivas es improductivo. Sería como tratar de definir la *materia* o el *espacio* en términos de algo más fundamental. Lo mejor que podemos hacer es dar ejemplos y caracterizaciones que se encuentren en el mismo nivel. Estas caracterizaciones no constituyen verdaderas definiciones debido a su naturaleza implícitamente circular, pero pueden ayudar a fijar aquello acerca de lo que se habla. Presupongo que todo lector tiene sus propias experiencias conscientes. Si todo resulta bien, estas caracterizaciones ayudarán a establecer que nuestro tema es justamente el de *aquellas* experiencias.

El tema quizá pueda caracterizarse mejor como "la cualidad subjetiva de la experiencia". Cuando percibimos, pensamos y actuamos, existe un ruido de fondo de causalidad y procesamiento de información, pero este procesamiento por lo general no ocurre en la oscuridad. Existe también un aspecto interno; hay algo que se siente como ser un agente cognitivo. Este aspecto interno es la experiencia consciente. Las experiencias conscientes van desde las vívidas sensaciones de colores hasta las experiencias de los más tenues aromas en el ambiente; desde agudos dolores a la huidiza experiencia de pensamientos en la punta de la lengua; desde sonidos y olores mundanos hasta la grandeza envolvente de la experiencia musical; desde la trivialidad de una fastidiosa comezón al peso de una profunda angustia existencial; desde la especificidad del sabor de la menta a la generalidad de la propia experiencia de uno mismo. Cada una de estas experiencias tiene una calidad experimentada distintiva. Todas son partes prominentes de la vida interior de la mente.

Podemos decir que un ser es consciente si existe *algo que es ser como* ese ser, para usar una frase que hizo famosa Thomas Nagel.[1] De forma similar, un estado mental es consciente si existe algo que es como estar en ese estado mental. Para plantearlo de otra manera, podemos decir que un estado mental es consciente si está ligado a una *sensación cualitativa*, una cualidad asociada de experiencia. Estas sensaciones cualitativas se conocen también como cualidades fenoménicas o *qualia* para abreviar.[2] El problema de explicar estas cualidades fenoménicas es justamente el problema de explicar la conciencia. Esta es la parte realmente difícil del problema mente-cuerpo.

¿Por qué debería existir la experiencia consciente? Es fundamental para un punto de vista subjetivo, pero desde un punto de vista objetivo es totalmente inesperada. Si adoptamos el punto de vista objetivo, podemos contar una historia acerca de cómo los campos, ondas y partículas en el continuo espaciotemporal interactúan de formas sutiles y llevan al desarrollo de sistemas complejos como el cerebro. En principio, no existe ningún profundo misterio filosófico en el hecho de que estos sistemas puedan procesar información de modos intrincados, reaccionar a estímulos con una conducta sofisticada e incluso exhibir capacidades complejas como el aprendizaje, la memoria y el lenguaje. Todo esto es impresionante, pero no es metafísicamente desconcertante. En contraste, la existencia de la experiencia consciente parece ser una *nueva* característica desde este punto de vista. No es algo que podríamos haber predicho a partir de las otras características solamente.

Esto es, la conciencia es *sorprendente*. Si todo lo que conocemos fuesen hechos de la física, o incluso hechos de la dinámica y el procesamiento de información en sistemas complejos, no habría ninguna razón apremiante para postular la existencia de la experiencia consciente. Si no fuese por nuestra evidencia directa como primera persona, la hipótesis podría parecer injustificada; casi mística, quizá. Sin embargo, sabemos de una forma directa que la experiencia consciente existe. La pregunta es, ¿cómo la reconciliamos con todo el resto de lo que sabemos?

La experiencia consciente es parte del mundo natural y como otros fenómenos naturales necesita desesperadamente una explicación. Existen aquí al menos dos objetivos principales de explicación. El primero y más fundamental es la propia *existencia* de la conciencia. ¿Por qué existe la experiencia consciente? Si surge en sistemas físicos, como parece probable, ¿cómo surge? Esto nos lleva a algunas preguntas más específicas. ¿La conciencia es ella misma física o es meramente un concomitante de sistemas físicos? ¿Cuán difundida está la conciencia? Los ratones, por ejemplo, ¿tienen experiencia consciente?

Un segundo objetivo es el *carácter* específico de las experiencias conscientes. Dado que la conciencia existe, ¿por qué las experiencias individuales tienen una naturaleza particular? Cuando abro los ojos y miro en torno de mi oficina, ¿por qué tengo *esta* clase de experiencia consciente? En un nivel más básico, ¿por qué ver el color rojo es *así*, y no *de otro modo*? Parece concebible que cuando miramos cosas rojas, como rosas rojas, podríamos tener la clase de experiencias de color que tenemos cuando miramos cosas azules. ¿Por qué es la experiencia de un modo y no de otro? ¿Por qué experimentamos una sensación

rojiza,[3] en lugar de alguna otra clase de sensación completamente diferente, como la del sonido de una trompeta?

Cuando alguien toca un do mayor en el piano, ocurre una compleja cadena de sucesos. El sonido vibra en el aire y una onda viaja hasta mi oído. La onda es procesada y analizada en frecuencias dentro del oído, y una señal es enviada a la corteza auditiva. Allí suceden nuevos procesamientos: el aislamiento de ciertos aspectos de la señal, su categorización y finalmente una reacción. Todo esto no es tan difícil de comprender en principio. Pero, ¿por qué debería estar acompañado por una *experiencia*? ¿Y por qué, en particular, debería estar acompañado por *esa* experiencia, con su característica riqueza tonal y de timbre? Estas son dos preguntas centrales que nos agradaría que una teoría de la conciencia respondiese.

Finalmente, nos gustaría que una teoría de la conciencia hiciese al menos lo siguiente: debería enunciar las condiciones bajo las cuales los procesos físicos dan origen a la conciencia y, para dichos procesos, debería especificar exactamente qué clases de experiencias están asociadas. Y nos gustaría que la teoría explicase *cómo* surge, de modo que la aparición de la conciencia nos parezca inteligible y no algo mágico. A la postre, nos gustaría que la teoría nos permita ver a la conciencia como una parte integral del mundo natural. En la actualidad resulta difícil poder vislumbrar cómo sería una teoría de esta clase, pero sin ella no podemos decir que comprendemos cabalmente la conciencia.

Antes de continuar, debemos hacer una nota sobre la terminología. El término "conciencia" es ambiguo, ya que refiere a una variedad de fenómenos distintos. A veces se utiliza para hacer referencia a una capacidad cognitiva, tal como la capacidad de hacer introspección o de informar sobre los propios estados mentales. A veces se utiliza como sinónimo de "vigilia". Otras veces está estrechamente ligado a nuestra capacidad de concentrar la atención o de controlar voluntariamente nuestra conducta. A veces "ser consciente de algo" se reduce a lo mismo que "saber acerca de algo". Todos estos son usos aceptados del término, pero todos ellos seleccionan fenómenos distintos del tema que estoy analizando, fenómenos que son significativamente menos difíciles de explicar. Más adelante volveré sobre estas nociones alternativas de la conciencia; por ahora, cuando hable de la conciencia, me referiré a la cualidad subjetiva de la experiencia: cómo es ser un agente cognitivo.

Un cierto número de términos y frases alternativas seleccionan aproximadamente la misma clase de fenómenos que la "conciencia" en su sentido central. Estos incluyen "experiencia", "qualia", "fenomenología", "fenoménico", "experiencia subjetiva" y "cómo es".

Aparte de las diferencias gramaticales, las diferencias entre estos términos representan cuestiones muy sutiles de connotación. "Ser consciente" en este sentido es aproximadamente sinónimo de "tener qualia", "tener experiencia subjetiva", etc. Cualquier diferencia en la clase de fenómenos seleccionados es insignificante. Como "conciencia", muchos de estos términos son algo ambiguos, pero nunca haré uso de estos términos en los sentidos alternativos. Utilizaré todas estas frases para hablar acerca del fenómeno central de este libro, pero "conciencia" y "experiencia" son los términos más directos y tenderán a reiterarse.

## Un catálogo de las experiencias conscientes

Es fascinante poner atención a la experiencia consciente. La experiencia viene en una enorme cantidad de variedades, cada una con sus propias características. En la siguiente lista impresionista y preteórica intentaré hacer un catálogo de los aspectos de la experiencia consciente que está muy lejos de ser completo. Nada aquí debería ser considerado demasiado seriamente como filosofía, pero debería ayudarnos a concentrar la atención en el tema de marras.

*Experiencias visuales*. Entre las muchas variedades de experiencia visual, las sensaciones de color se destacan como ejemplos paradigmáticos de la experiencia consciente debido a su pura y aparentemente inefable naturaleza cualitativa. Algunas experiencias de color parecen sorprendentes, por lo que pueden ser particularmente buenas para concentrar nuestra atención sobre el misterio de la conciencia. En este momento en mi ambiente hay un matiz particularmente intenso de púrpura proveniente de un libro colocado en un estante; un tono casi surrealista de verde en una fotografía de helechos sobre una pared; un conjunto centelleante de luces rojo brillante, verde, naranja y azul sobre un árbol de Navidad que puedo ver a través de la ventana. Pero cualquier color puede producirnos asombro si le prestamos atención y reflexionamos sobre su naturaleza. ¿Por qué debería experimentarse *así*? ¿Por qué debería experimentarse de alguna forma? ¿Cómo puedo transmitir la naturaleza de esta experiencia de color a alguien que no la haya tenido?

Otros aspectos de la experiencia visual incluyen la experiencia de la forma, el tamaño, el brillo y la oscuridad. Un aspecto particularmente sutil es la experiencia de la profundidad. Cuando era niño, uno de mis ojos tenía una visión excelente, pero la del otro era muy pobre. Gracias a mi ojo bueno, el mundo parecía nítido y punzante, y ciertamente tridimensional. Un día, comencé a usar anteojos y el cambio fue notable. El mundo no fue mucho más punzante que antes,

pero de pronto me pareció *más* tridimensional: cosas que antes tenían profundidad ahora parecían más profundas, y el mundo se había vuelto un lugar más opulento. Si usted se cubre un ojo y luego lo descubre, podrá tener una idea del cambio. En mi estado anterior, yo habría dicho que no había ningún modo como la profundidad de mi visión pudiese mejorar; el mundo ya parecía todo lo tridimensional que podía ser. El cambio fue sutil, casi inefable, pero extremadamente sorprendente. Por supuesto, existe una historia intelectual que podemos contar acerca de cómo la visión binocular permite que la información de cada ojo se consolide en información acerca de distancias, lo que hace posible un control más sofisticado de la acción, pero de algún modo esa historia causal no revela el modo en el que *sentía* la experiencia. Por qué ese cambio en el procesamiento debería estar acompañado por una modificación semejante de mi experiencia me resultó misterioso cuando tenía diez años de edad y todavía es, para mí, una fuente de admiración.

*Experiencias auditivas*. En cierto modo, los sonidos son aun más extraños que las imágenes visuales. La estructura de las imágenes por lo general corresponde directamente a la estructura del mundo, pero los sonidos pueden ser bastante independientes. Entra una llamada a mi teléfono, un dispositivo interno vibra, se inicia una onda compleja en el aire y luego llega a mis tímpanos, y yo, de algún modo, casi mágicamente, escucho un *campanilleo*. Ninguna cualidad del campanilleo parece corresponder directamente a alguna estructura en el mundo, aunque sé que se originó en el aparato telefónico y que está determinado por una forma de onda. ¿Por qué debería esa forma de onda, o incluso los disparos neuronales involucrados, dar origen a una cualidad sonora como *esa*?

La experiencia musical es quizás el aspecto más rico de la experiencia auditiva, aunque la experiencia del habla debería estar cercana. La música es capaz de transportarnos y absorbernos por completo, envolviéndonos del modo como un campo visual puede hacerlo, pero que las experiencias auditivas usualmente no lo hacen. Podemos examinar los aspectos de la experiencia musical analizando los sonidos que percibimos en notas y tonos con complejas interrelaciones entre sí, pero la experiencia de la música de alguna forma va más allá de esto. Una experiencia cualitativa unificada surge de un acorde, pero no de notas seleccionadas al azar. Un antiguo piano y un oboe distante pueden combinarse para producir una experiencia inesperadamente conmovedora. Como siempre, cuando reflexionamos sobre la cuestión, nos hacemos la pregunta: ¿por qué debería *eso* sentirse como *esto*?




*Figura 1.1.* Efabilidad e inefabilidad en la experiencia olfativa. (Calvin y Hobbes © Watterson. Distribuido por Universal Press Syndicate. Reproducido con autorización. Todos los derechos reservados.)

*Experiencias táctiles.* Las texturas proporcionan otro de los espacios más ricos en cualidades que experimentamos: piense en la sensación del terciopelo y contrástela con la textura del frío metal, o una mano pegajosa o un mentón con una barba incipiente. Todos estos tienen su propia y única cualidad. Las experiencias táctiles del agua, el algodón de azúcar o de los labios de otra persona son, reitero, diferentes.

*Experiencias olfativas.* Piense en el olor mohoso de un viejo guardarropas, el hedor de basura pudriéndose, la fragancia del césped recién cortado, el tibio aroma del pan recién horneado. El olfato es, en cierta forma, el más misterioso de todos los sentidos, debido a la naturaleza rica, intangible e indescriptible de las sensaciones olfativas. Ackermann (1990) lo llama "el sentido mudo; el que no tiene palabras". Aunque existe algo inefable en cualquier sensación, los otros sentidos tienen propiedades que facilitan su descripción. Las experiencias visuales y auditivas tienen una estructura combinatoria compleja que puede describirse. Las experiencias táctiles y gustativas por lo general surgen del contacto directo con algún objeto, y existe un rico vocabulario descriptivo que hace referencia a esos objetos. El olfato tiene poco en lo que respecta a una estructura evidente y con frecuencia flota libre de cualquier objeto manifiesto, por lo que resulta una presencia primitiva en nuestra variedad sensorial (Quizá los animales puedan tener un mejor desempeño [fig. 1.1]). El carácter primitivo quizá se deba en parte al proceso de "encaje" por el cual nuestros receptores olfativos son sensibles a diversos tipos de moléculas. Resulta arbitrario que un tipo dado de molécula origine *este* tipo de sensación, pero así lo hace.

*Experiencias gustativas*. La investigación psicofísica nos dice que sólo hay cuatro dimensiones independientes en la percepción del gusto: dulce, ácido, amargo y salado. Pero este espacio cuatridimensional se combina con nuestro sentido olfativo para producir una gran variedad de experiencias posibles: el sabor de la delicia turca, de la ensalada de arvejas al curry,[4] de una pastilla de menta, de un durazno maduro.

*Experiencias de frío y calor*. Un día opresivamente tórrido y húmedo y un helado día de invierno producen experiencias cualitativas sorprendentemente diferentes. Piense también en las sensaciones de calor sobre la piel cuando estamos cerca de un fuego y en la sensación de frío quemante que obtenemos al tocar hielo seco.

*Dolor*. El dolor es un ejemplo paradigmático de experiencia consciente amado por los filósofos. Quizás, esto se deba a que los dolores forman una clase muy distintiva de experiencias cualitativas, y es difícil ponerlas en correspondencia directa con alguna estructura en el mundo o en el cuerpo, aunque usualmente están asociadas con alguna parte del cuerpo. Debido a esta característica, los dolores pueden parecer aun más subjetivos que la mayoría de las experiencias sensoriales. Existe una gran variedad de experiencias de dolor, desde dolores punzantes y fieras quemaduras pasando por agudas punzadas hasta dolores apagados.

*Otras sensaciones corporales*. Los dolores sólo son la clase más notable de sensaciones asociadas con partes determinadas del cuerpo. Otras sensaciones son las cefaleas (que quizá sean una clase de dolor), los retortijones producidos por el hambre, la picazón, el cosquilleo y la experiencia asociada con la necesidad de orinar. Muchas sensaciones corporales tienen una cualidad completamente única, de una clase diferente de cualquier otra de nuestras experiencias: piénsese en los orgasmos, o la sensación de golpearse el nervio en el codo. También existen experiencias asociadas con la propiocepción, el sentido de dónde está el propio cuerpo en el espacio.

*Imaginería mental*. Moviéndonos cada vez más adentro, hacia experiencias que no están asociadas con objetos particulares en el ambiente o el cuerpo sino que, en algún sentido, se generan internamente, llegamos a las imágenes mentales. Existe una rica fenomenología asociada con las imágenes visuales evocadas por la imaginación, aunque estas no son tan detalladas como las que surgen de la

percepción visual directa. También existen interesantes patrones de colores que se obtienen cuando entornamos los ojos, y las intensas imágenes posteriores que se producen luego de mirar algo brillante. Nuestra imaginación puede evocar también clases similares de "imágenes" auditivas, e incluso imágenes táctiles, olfativas y gustativas, aunque estas son más difíciles de fijar y la sensación cualitativa asociada es por lo general más débil.

*Pensamiento consciente*. Algunas de las cosas que pensamos y creemos no tienen asociadas ninguna sensación cualitativa particular, pero muchas sí. Esto se aplica especialmente a nuestros propios pensamientos explícitos y a diversos pensamientos que afectan al propio flujo de la conciencia. Con frecuencia es difícil determinar en forma precisa cuál es la sensación cualitativa de un pensamiento, pero seguramente allí está. Hay *algo* que es como tener dichos pensamientos.

Cuando pienso en un león, por ejemplo, parece haber un hálito de cualidad leonina en mi fenomenología: pensar en un león es sutilmente diferente de pensar en la torre Eiffel. En términos más evidentes, las actitudes cognitivas como el deseo suelen tener una intensa sensación fenoménica. El deseo parece ejercer un "tirón" fenomenológico, y la memoria suele tener un componente cualitativo, como en la experiencia de nostalgia o pesar.

*Emociones*. Las emociones suelen estar asociadas con experiencias distintivas. La chispa de un estado de ánimo alegre, el recelo de una profunda depresión, el fulgor rojo vivo de un arranque de ira, la melancolía del pesar: todos ellos pueden afectar profundamente la experiencia consciente, aunque de un modo mucho menos específico que experiencias localizadas como las sensaciones. Las emociones impregnan y colorean todas nuestras experiencias conscientes mientras duran.

Otros sentimientos más transitorios se encuentran a mitad de camino entre las emociones y los aspectos más obviamente cognitivos de la mente. Piense en la oleada de placer que sentimos cuando escuchamos un chiste. Otro ejemplo es la sensación de tensión que experimentamos cuando observamos una película de suspenso, o cuando aguardamos un suceso importante. El cosquilleo en la boca del estómago que puede acompañar al nerviosismo también cae en esta clase.

*El sentido del sí mismo*. A veces sentimos que hay algo en la experiencia consciente que trasciende a todos estos elementos espe-

cíficos: una especie de ruido de fondo, por ejemplo, que es de algún modo fundamental para la conciencia y que está allí incluso cuando otros componentes no lo están. Esta fenomenología del sí mismo es tan profunda e intangible que a veces parece ilusoria, que consiste nada más que en elementos específicos como los nombrados antes. Sin embargo, parece haber *algo* en la fenomenología del sí mismo, aun cuando sea muy difícil de determinar.

Este catálogo cubre un número de áreas, pero deja afuera tanto como incluye. No dije nada, por ejemplo, de los sueños, la excitación y la fatiga, la embriaguez, o el carácter novedoso de otras experiencias inducidas por las drogas. También existen ricas experiencias que derivan sus características de la combinación de dos o más de los componentes descriptos más arriba. Mencioné los efectos combinados del olfato y el gusto, pero un ejemplo igualmente sobresaliente es la experiencia combinada de la música y la emoción, que interactúan de un modo sutil, difícil de separar. También dejé de lado la unidad de la experiencia consciente, el modo como todas estas experiencias parecen estar vinculadas entre sí como las experiencias de un solo experimentador. Al igual que el sentido del sí mismo, esta unidad a veces parece ilusoria —es ciertamente más difícil de captar que cualquier experiencia específica— pero tenemos una intensa intuición de que la unidad existe.

Lamentablemente, no volveremos a ocuparnos de la rica variedad de la experiencia consciente. En el estudio de los misterios filosóficos asociados con esta, una simple sensación de color plantea los problemas tan profundamente como la experiencia de una coral de Bach. Las cuestiones más importantes se encuentran en todas estas variedades, de manera que la consideración de la naturaleza de experiencias específicas no es particularmente relevante. Con todo, este breve examen de la rica variedad de la experiencia consciente debería ayudar a concentrar nuestra atención en exactamente aquello que está en discusión, y proporcionar un surtido de ejemplos que pueden tenerse en mente durante el análisis más abstracto.[5]

## 2. Los conceptos fenoménico y psicológico de la mente

La experiencia consciente no es lo único que hay en la mente. Para darse cuenta de ello, obsérvese que aunque la ciencia cognitiva moderna no tuvo casi nada que decir acerca de la conciencia, sí tuvo mucho que decir sobre la mente en general. Los aspectos de la mente de los que se ocupa son diferentes. La ciencia cognitiva intenta

fundamentalmente explicar la conducta, y se ocupa de la mente en tanto la interpreta como la base interna de la conducta, y de los estados mentales en tanto los interpreta como aquellos estados relevantes a la causalidad y la explicación de la conducta. Pero los estados pueden o no ser conscientes. Desde el punto de vista de la ciencia cognitiva, un estado interno responsable de las causas de la conducta es mental independientemente de que sea o no consciente.

En el origen de todo esto se encuentran dos conceptos distintos de la mente. El primero es el *fenoménico*. Este es el concepto de la mente como experiencia consciente y el de un estado mental como estado mental conscientemente experimentado. Se trata del aspecto más desconcertante de la mente y en el que me concentraré, pero esto no agota lo mental. El segundo es el concepto *psicológico*, el concepto de la mente como base causal o explicativa de la conducta. Un estado es mental en este sentido si desempeña el papel causal apropiado en la explicación de la conducta. Según el concepto psicológico, importa poco si un estado mental tiene una cualidad consciente o no. Lo que importa es el papel que desempeña en una economía cognitiva.

En el concepto fenoménico, la mente se caracteriza por el modo como se la *experimenta*; en el concepto psicológico, la mente se caracteriza por lo que *hace*. No debería haber ninguna cuestión de competencia entre estas dos nociones. Ninguna de las dos es *el* análisis correcto de la mente. Cubren diferentes fenómenos y ambas son muy reales.

A veces hablaré de los "aspectos" fenoménicos y psicológicos de la mente, y a veces de la "mente fenoménica" y la "mente psicológica". En esta etapa inicial, no deseo formular ninguna presuposición acerca de si lo fenoménico y lo psicológico pueden ser la misma cosa. Quizá todo estado fenoménico sea un estado psicológico, en el sentido de que posee un papel significativo en las causas y la explicación de la conducta, y quizá todo estado psicológico tenga una relación íntima con lo fenoménico. Por ahora, lo importante es la distinción conceptual entre las dos nociones: que un estado sea fenoménico *significa* que se experimenta de cierto modo, y que un estado sea psicológico significa que desempeña un papel causal apropiado. Estas nociones diferentes no deben mezclarse, al menos al comienzo.

Un concepto mental específico puede usualmente analizarse como un concepto fenoménico, como un concepto psicológico o como una combinación de los dos. Por ejemplo, la sensación, en su sentido central, es mejor interpretarla como un concepto fenoménico: tener una sensación es tener un estado con un cierto tipo de sensación. Por otro lado, conceptos como aprendizaje y memoria es mejor interpre-

tarlos como conceptos psicológicos. Que alguien aprenda es, en primera aproximación, que adapte sus capacidades conductuales de un modo apropiado en respuesta a ciertos tipos de estimulación ambiental. En general, una propiedad fenoménica de la mente se caracteriza por cómo es para un sujeto experimentar esa propiedad, mientras que una propiedad psicológica se caracteriza por su papel asociado en la causalidad y/o explicación de la conducta.

Por supuesto, esta utilización del término "psicológico" es una estipulación: surge de identificar la psicología con la ciencia cognitiva tal como se describió más arriba. El concepto cotidiano de "estado psicológico" es probablemente más amplio que esto, y puede bien incluir elementos de lo fenoménico. Pero nada importante dependerá de mi utilización del término.

## Una historia enlatada

Los aspectos fenoménico y psicológico de la mente tienen una larga historia de fusión. Es probable que René Descartes haya sido parcialmente responsable de esto. Con su notoria doctrina de que la mente es transparente para sí misma, se acercó a identificar lo mental con lo fenoménico. Descartes sostenía que todo suceso en la mente es una *cogitatio*, o un contenido de la experiencia. A esta clase asimilaba las voliciones, las intenciones y cualquier otro tipo de pensamiento. En su respuesta al Cuarto Conjunto de Objeciones, escribió:

> En cuanto al hecho de que no puede haber nada en la mente, en tanto ente pensante, de lo que no se haya percatado, esto me parece obvio. Porque no hay nada que podamos suponer que está en la mente, considerada de este modo, que no sea un pensamiento o dependa de un pensamiento. Si no fuese un pensamiento ni dependiese de un pensamiento no pertenecería a la mente *en tanto* ente pensante; y no podemos tener ningún pensamiento del que no estemos conscientes cuando este se encuentra en nosotros.

Aunque Descartes no identificó concretamente lo psicológico con lo fenoménico, al menos supuso que todo lo psicológico que merece ser llamado mental tiene un aspecto consciente.[6] Para Descartes, la noción de estado mental inconsciente era una contradicción.

Los progresos en la teoría psicológica, no en la filosofía, fueron responsables de distinguir los dos aspectos de la mente. Hace tan sólo un siglo, psicólogos como Wilhelm Wundt y William James eran manifiestamente cartesianos en el sentido de que usaban la intros-

pección para investigar las causas de la conducta, y desarrollaron teorías psicológicas sobre la base de evidencia introspectiva. De esta manera, la fenomenología se transformó en el árbitro de la psicología. Pero los desarrollos inmediatamente posteriores instauraron al psicológico como un dominio autónomo.

Especialmente Sigmund Freud y sus contemporáneos consolidaron la idea de que muchas actividades de la mente son inconscientes y que puede haber cosas como creencias y deseos inconscientes. El hecho mismo de que esta noción parecía coherente evidencia que se estaba utilizando un análisis no fenomenológico del pensamiento. Al parecer Freud interpretó las nociones *causalmente*. El deseo, en líneas generales, era implícitamente interpretado como la clase de estado que causa un cierto tipo de conducta asociada con el objeto del deseo. La creencia se interpretaba según su papel causal de un modo similar. Por supuesto, Freud no hizo estos análisis explícitos, pero algo similar subyace claramente en su utilización de las nociones. Reconoció de un modo explícito que la accesibilidad a la conciencia no es esencial para la relevancia de un estado en la explicación de la conducta, y que una cualidad consciente no es esencial para que algo sea una creencia o un deseo. Estas conclusiones se apoyan en una noción de lo mental que es independiente de las nociones fenoménicas.

Alrededor de la misma época, el movimiento conductista en psicología había rechazado totalmente la tradición introspeccionista. Se desarrolló una nueva clase "objetiva" de explicación, sin ningún lugar para la conciencia en sus explicaciones. Esta forma de explicación sólo tuvo un éxito parcial, pero consolidó la idea de que la explicación psicológica puede progresar ignorando lo fenoménico. Los conductistas diferían en sus posiciones teóricas: algunos reconocían la existencia de la conciencia pero la encontraban irrelevante para la explicación psicológica y algunos directamente negaban su existencia. Muchos iban todavía más lejos y negaban la existencia de *cualquier* clase de estado mental. La razón oficial de esto era que se suponía que los estados internos eran irrelevantes en la explicación de la conducta y que esta podía realizarse enteramente en términos externos. Una razón más profunda es, quizá, que todas las nociones mentales estaban teñidas por el desprestigiado olor de lo fenoménico.

En cualquier caso, estos dos desarrollos establecieron como ortodoxia la idea de que la explicación de la conducta no depende de ningún modo de nociones fenoménicas. El pasaje del conductismo a la ciencia cognitiva computacional preservó en buena medida esa ortodoxia. Aunque este último movimiento recuperó un papel para los estados internos, que podían incluso llamarse estados "mentales", no había nada particularmente fenoménico en ellos. Esos estados eran

admisibles precisamente sobre la base de su relevancia para la explicación de la conducta; cualquier cualidad fenoménica asociada era, cuanto más, irrelevante. El concepto de lo mental como psicológico ocupó, de esta manera, el centro de la escena.

En filosofía, el traslado del interés de lo fenoménico a lo psicológico fue codificado por Gilbert Ryle (1949), quien argumentó que todos nuestros conceptos mentales pueden analizarse en términos de ciertas clases de conductas asociadas, o en términos de disposiciones a comportarse de ciertos modos.[7] Este enfoque, el conductismo lógico, es manifiestamente el precursor de gran parte de lo que se considera la ortodoxia en la filosofía de la psicología contemporánea. En particular, fue la codificación más explícita del vínculo entre los conceptos mentales y la causalidad de la conducta.

Ryle no formuló esta teoría como un análisis de sólo *algunos* conceptos mentales. Su intención era que todos ellos cayesen dentro de su alcance. A muchos les pareció, como me parece a mí, que, en tanto análisis de nuestros conceptos fenoménicos, tales como la sensación y la conciencia, este enfoque es un fracaso. A muchas personas les resultaba evidente que cuando hablamos sobre estados fenoménicos, no hablamos de nuestra conducta, o de alguna disposición conductual. Pero, de cualquier forma, el análisis de Ryle proporcionó un enfoque sugerente de muchas otras nociones mentales, como creer, disfrutar, querer, pretender y recordar.

Aparte de sus problemas con los estados fenoménicos, el enfoque de Ryle adolecía de algunas dificultades técnicas. Primero, es natural suponer que los estados mentales *causan* la conducta, pero si los estados mentales son ellos mismos conductas o disposiciones conductuales, en lugar de estados internos, es difícil ver cómo podrían hacer el trabajo. Segundo, algunos argumentaron (Chisholm [1957] y Geach [1957]) que ningún estado mental puede definirse mediante un único rango de disposiciones conductuales, independientemente de otros estados mentales. Por ejemplo, si creemos que está lloviendo, nuestras disposiciones conductuales variarán según cuál sea nuestro deseo de mojarnos. Por lo tanto, es necesario invocar otros estados mentales para caracterizar las disposiciones conductuales asociadas a una clase determinada de estados mentales.

Estos problemas fueron refinados por lo que llegó a conocerse como *funcionalismo*, desarrollado por David Lewis (1966) y con más profundidad por David Armstrong (1968).[8] Según este enfoque, un estado mental se define cabalmente por su *papel causal*: esto es, en términos de las clases de estimulación que tienden a producirlo, el tipo de conducta que tiende a producir y el modo como interactúa con otros estados mentales. Este enfoque consideraba que los estados

mentales son totalmente internos y capaces de guardar la clase correcta de relación causal con la conducta, respondiendo así a la primera objeción, y hacía posible que los estados mentales se definiesen en términos de su interacción, respondiendo así a la segunda objeción.

Desde este punto de vista, nuestros conceptos mentales pueden analizarse *funcionalmente*: en términos de sus causas y efectos concretos o típicos. Hacer un análisis de esta clase para cualquier concepto mental determinado es muy poco trivial; Armstrong (1968) realiza algunos análisis de este tipo, pero son muy incompletos. Como posición teórica, sin embargo, el funcionalismo puede proporcionar una interpretación razonable de muchos de nuestros conceptos mentales, al menos en la medida en que tienen un papel en la explicación de la conducta. Por ejemplo, la noción de aprendizaje podría analizarse como la adaptación de las propias capacidades conductuales en respuesta a la estimulación ambiental. Para considerar un estado más complejo, una creencia de que está lloviendo podría analizarse a grandes rasgos como la clase de estado que tiende a producirse cuando está lloviendo, y que lleva a conductas que serían apropiadas si estuviese lloviendo, que interactúa inferencialmente de algún modo con otras creencias y deseos, etc. Hay amplio espacio para resolver los detalles, pero muchos pensaron que la idea general estaba en el camino correcto.

Como Ryle, sin embargo, Armstrong y Lewis no realizaron esta formulación como un análisis de *algunos* conceptos mentales. Suponían, en cambio, que se trataba de un análisis de todos los conceptos mentales. En particular, argumentaban que las nociones de experiencia, sensación, conciencia, etc., podían analizarse de esta forma. Esta asimilación de lo fenoménico a lo psicológico me impresiona como un error de la misma magnitud que la asimilación de Descartes de lo psicológico a lo fenoménico. Simplemente es un falso análisis de lo que significa ser fenoménico. Cuando nos preguntamos si alguien tiene una experiencia de color, no nos preguntamos si está recibiendo estimulación ambiental y procesándola de cierto modo. Nos preguntamos si está *experimentando* una sensación de color, y esta es una pregunta distinta. Es una posibilidad conceptualmente coherente de que algo pueda desempeñar un papel causal sin que exista una experiencia asociada.

Para decirlo de un modo diferente, nótese que este análisis de los conceptos fenoménicos no esclarece por qué alguien puede haberse ocupado alguna vez de este problema.[9] No es ningún gran misterio que un estado pueda tener un papel causal, aunque seguramente en ello hay problemas técnicos para la ciencia. Lo que sí resulta misterioso es por qué ese estado debería *experimentarse* como algo; por qué

debería tener una cualidad fenoménica. Son dos cuestiones totalmente diferentes por qué un estado desempeña un papel causal y por qué posee una cualidad fenoménica. El análisis funcionalista niega que estas preguntas sean diferentes y por lo tanto resulta insatisfactorio.

Consideraré esta cuestión con mucho mayor detalle más adelante, pero por ahora podemos notar que aunque la concepción funcionalista produzca un análisis insatisfactorio de los conceptos fenoménicos, podría proporcionar un buen análisis de otras nociones mentales, como el aprendizaje, la memoria y quizá las creencias. No surgen objeciones paralelas en estos casos. No resulta más misterioso que un sistema sea capaz de aprender, que el hecho de que un sistema sea capaz de adaptar su conducta en respuesta a la estimulación ambiental; ciertamente, parecen ser más o menos la misma cuestión. De un modo similar, cuando nos preguntamos si alguien aprendió algo, parece razonable decir que al hacerlo nos estamos preguntando si habrá sufrido un cambio que producirá en él una mejor capacidad para enfrentar ciertas situaciones en el futuro. Por supuesto, un análisis completo del concepto de aprendizaje será más sutil que esta primera aproximación, pero los demás detalles del análisis podrán especificarse dentro de este mismo marco general.

La concepción funcionalista corresponde precisamente a la definición que enuncié de las propiedades *psicológicas*. La mayoría de las propiedades mentales no fenoménicas caen dentro de esta clase y pueden, por lo tanto, ser analizadas de un modo funcional. Seguramente hay espacio suficiente para debatir acerca de los detalles de un análisis funcionalista específico. También existen preguntas teóricas significativas acerca de cuestiones como el papel del ambiente en la caracterización de las propiedades psicológicas, y si la causalidad, la explicación o ambas proporcionan el vínculo definitorio ente las propiedades psicológicas y la conducta. Sin embargo, estos detalles son relativamente poco importantes aquí. Lo que sí es importante es que los estados mentales no fenoménicos se caracterizan principalmente por su papel en nuestra economía cognitiva.

La moraleja de esta exposición es que lo psicológico y lo fenoménico son aspectos reales y distintos de la mente. En una primera aproximación, los conceptos fenoménicos tratan de los aspectos de primera persona de la mente, y los conceptos psicológicos de los aspectos de tercera persona. Nuestra perspectiva hacia la mente será bastante diferente según en qué aspectos de ella estemos interesados. Si lo que nos interesa es el papel de la mente para causar la conducta, nos concentraremos en las propiedades psicológicas. Si lo que nos interesa es la experiencia consciente de los estados mentales, nos concentraremos en las propiedades fenoménicas. Ni lo fenoménico ni lo

psicológico debería definirse eliminativamente en términos del otro. Es concebible que algún análisis profundo pueda revelar un vínculo fundamental entre lo fenoménico y lo psicológico, pero esto sería una tarea no trivial, y no es algo que pueda realizarse por estipulación previa. Asimilar lo fenoménico a lo psicológico previamente a alguna explicación profunda sería restar importancia al problema de la experiencia consciente; y asimilar lo psicológico a lo fenoménico sería limitar vastamente el papel de lo mental en la explicación de la conducta.

## 3. La doble vida de los términos mentales

Parece razonable decir que en conjunto lo psicológico y lo fenoménico agotan lo mental. Esto es, toda propiedad mental es una propiedad fenoménica, una propiedad psicológica, o una combinación de las dos. Ciertamente, si nos ocupamos de aquellas propiedades manifiestas de la mente que reclaman una explicación, encontramos, primero, las variedades de la experiencia consciente y, segundo, la causalidad de la conducta. No existe ningún tercer tipo de explanandum manifiesto, y las primeras dos fuentes de evidencia —la experiencia y la conducta— no nos dan ningún motivo para creer en la existencia de alguna tercera clase de propiedades no funcionales y no fenoménicas (quizá con una excepción menor para las propiedades relacionales que discutiremos en breve). Existen otras clases de estados mentales de los que hablamos con frecuencia —estados intencionales, estados emocionales, etc.— pero es plausible que estos puedan ser asimilados a lo psicológico, a lo fenoménico o a una combinación de los dos.

Las cosas se complican por el hecho de que muchos conceptos mentales cotidianos se encuentran en el medio: tienen un componente fenoménico y uno psicológico. El *dolor* suministra un claro ejemplo. Este término se utiliza frecuentemente para denominar un tipo particular de cualidad fenoménica displacentera, por lo que una noción fenoménica es central. Pero también existe una noción psicológica asociada con el término: aproximadamente, el concepto del tipo de estado que tiende a ser producido por algún daño al organismo, tiende a que se produzcan reacciones de aversión, etc. Ambos aspectos son fundamentales para la noción de sentido común del dolor. Podríamos decir que la noción de dolor es ambigua entre el concepto fenoménico y el concepto psicológico, o podríamos decir que ambos son componentes de un único concepto complejo.

Podemos complicarnos con todo tipo de nudos conceptuales si nos preocupamos acerca de si la cualidad fenoménica o el papel funcional es más esencial en el dolor. Por ejemplo, ¿un sistema

hipotético en el que se satisficieran todos los criterios funcionales pero que no tuviese una experiencia consciente, estaría verdaderamente dolorido? Podríamos sentirnos tentados a decir que no, pero ¿qué hay del hecho de que hablemos de dolores que duran un día, aunque no seamos conscientes de ellos en todo momento? Tiene poco sentido tratar de legislar sobre la cuestión en un sentido o en otro. Nada importante depende de la decisión semántica de si alguna cualidad fenoménica es *realmente* esencial para que algo pueda ser considerado un dolor. Podemos, en cambio, reconocer los diferentes componentes asociados con un concepto y distinguirlos explícitamente, hablando, por ejemplo, de "dolor fenoménico" o "dolor psicológico". Nuestro concepto cotidiano de dolor combina posiblemente los dos aspectos en una sutil mezcla ponderada, pero, para la discusión filosófica, las cosas serán más claras si las mantenemos separadas.

La razón de por qué las propiedades fenoménicas y psicológicas suelen utilizarse en forma conjunta es clara: se debe a que las propiedades relevantes tienden a coocurrir. Por lo general, cuando se desarrollan procesos que son resultado de un daño de tejidos y llevan a una reacción de aversión, se instancia algún tipo de cualidad fenoménica. Esto es, cuando el dolor psicológico está presente, el dolor fenoménico por lo general también está presente. No es una verdad *conceptual* que el proceso deba estar acompañado por la cualidad fenoménica, es un hecho acerca del mundo. Dada esta especie de coocurrencia de propiedades en las situaciones diarias, es natural que nuestros conceptos cotidianos las aúnen.

Muchos conceptos mentales llevan este tipo de doble vida Por ejemplo, el concepto de *percepción* puede interpretarse de una forma exclusivamente psicológica como el proceso por el cual los sistemas cognitivos son sensibles a la estimulación ambiental de una manera tal que los estados resultantes desempeñan un cierto papel en la dirección de los procesos cognitivos. Pero también se la puede interpretar fenoménicamente como la experiencia consciente de lo que se percibe. La posibilidad de la experiencia consciente subliminal va en contra de la última interpretación, pero algunos argumentarían que esta puede calificarse como percepción sólo en un sentido debilitado del término. Una vez más, sin embargo, la cuestión es terminológica. Cuando queremos ser claros, podemos simplemente estipular si nos estamos ocupando de la propiedad psicológica, de la propiedad fenoménica o de una combinación de ambas.

Algunos de estos conceptos duales, sin embargo, se inclinan más hacia lo fenoménico y otros se inclinan más hacia lo psicológico. Tómese el concepto de *sensación*, que está estrechamente relacionado con el concepto de percepción y que tiene también componentes

fenoménicos y psicológicos. El componente fenoménico es mucho más prominente en "sensación" que en "percepción", como lo atestigua el hecho de que la idea de percepción inconsciente parece tener más sentido que la de sensación inconsciente. Las cosas son todavía algo grises —queda un sentido de "percepción" que requiere de la experiencia consciente, y un sentido de "sensación" que no— pero estos sentidos parecen menos centrales que las alternativas. Quizá sea más natural utilizar "percepción" como término psicológico, y "sensación" como término fenoménico. De este modo, podemos interpretar la sensación como algo similar a la contraparte fenoménica de la percepción.

Un buena prueba para determinar si una noción mental M es principalmente psicológica es preguntarse: ¿Podría algo ser una instancia de M sin ninguna cualidad fenoménica asociada? Si así es, entonces es probable que M sea psicológico. Si no, entonces M es fenoménico, o al menos una noción combinada que involucra de un modo fundamental a la fenomenología. No podemos descartar esta última posibilidad, ya que algunos conceptos requieren una clase apropiada de cualidad fenoménica y un papel cognitivo adecuado; por ejemplo, es posible que un sentido central de "sensación" tenga este carácter combinado. Pero, podemos, al menos, separar las nociones que involucran a la fenomenología de aquellas que no lo hacen.

La prueba sugiere que un concepto como *aprendizaje*, por ejemplo, es fundamentalmente psicológico. En una primera aproximación, el aprendizaje consiste en que las propias capacidades se adapten de cierta manera a diversas circunstancias y estímulos nuevos. No se requiere ninguna cualidad fenoménica particular para que un proceso cognitivo sea una instancia de aprendizaje; estas cualidades pueden estar presente, pero no es lo que hace que el proceso sea un ejemplo de aprendizaje. Podría haber un tenue matiz fenoménico heredado de un vínculo con conceptos como el de creencia, que analizaremos más abajo, pero este es sumamente débil. Al explicar el aprendizaje, lo fundamental que debemos aclarar es cómo el sistema consigue adaptarse del modo apropiado. Algo similar ocurre con conceptos como los de categorización y memoria, que parecen ser esencialmente nociones psicológicas, en el sentido de que lo fundamental es el desempeño de un papel cognitivo apropiado.

Las *emociones* tienen un aspecto fenoménico mucho más claro. Cuando pensamos en la felicidad o la tristeza, nos vienen a la mente variedades distintas de la experiencia consciente. Sin embargo, no es del todo evidente que el aspecto fenoménico sea esencial para que un estado sea una emoción; sin duda también existe una fuerte propiedad psicológica asociada. Como es usual, no estamos obligados a

tomar ninguna decisión sobre esta cuestión. Simplemente podemos hablar sobre los aspectos psicológicos y fenoménicos de la emoción, y observar que estos agotan los aspectos de la emoción que requieren explicación.

El caso más complejo es el de los estados mentales como la *creencia*, frecuentemente denominados "actitudes proposicionales" porque son actitudes hacia las proposiciones acerca del mundo. Cuando *creo* que Bob Dylan hará una gira por Australia, por ejemplo, doy crédito a una cierta proposición acerca de Dylan; cuando *deseo* que Dylan haga una gira por Australia, tengo una actitud diferente hacia la misma proposición. La característica central de estos estados mentales es su aspecto semántico, o *intencionalidad*: el hecho de que son todos *acerca de* cosas en el mundo. Es decir, una creencia tiene contenido semántico: el contenido de mi creencia citada más arriba es algo así como la proposición de que Dylan hará una gira por Australia (aunque hay espacio para el debate aquí).

La creencia suele considerarse una propiedad psicológica. Según este enfoque, en primera aproximación, creer que una proposición es verdadera es estar en un estado en el que actuamos de un modo que sería apropiado si fuese verdad, un estado que tiende a producirse por ser el caso, y un estado en el cual nuestra propia dinámica cognitiva del razonamiento refleja la interacción apropiada de la creencia con otras creencias y deseos. Los criterios funcionales de las creencias son muy sutiles, sin embargo, y nadie produjo hasta ahora nada que se parezca a un análisis completo de los criterios relevantes. A pesar de todo, hay razones para creer que este enfoque captura buena parte de lo que es significativo acerca de las creencias. Se relaciona con la idea de que la creencia es una especie de *construcción explicativa*: atribuimos creencias a otros principalmente para explicar su conducta.

Algunos argumentarían que este análisis es incompleto, y que se requiere algo más, además del tipo relevante de proceso psicológico, para que algo sea una creencia. En particular, deja afuera los aspectos *experienciales* de creer, que, según algunos, son esenciales para que algo cuente como una creencia. Por ejemplo, Searle (1990a) argumentó que el contenido intencional de una creencia depende enteramente del estado asociado de la conciencia, o de un estado de la conciencia que la creencia puede originar. Sin conciencia, todo lo que existe es una intencionalidad "como si".[10]

Por cierto, a menudo existen experiencias conscientes relacionadas con la creencia: hay algo que es como tener una creencia acaeciente (esto es consciente), y la mayoría de las creencias no acaecientes pueden al menos producir una creencia consciente. Las cuestiones cruciales son, sin embargo, si esa cualidad consciente es lo

que *hace* que el estado sea una creencia, y si es lo que le da el contenido que posee. Esto puede ser más plausible para algunas creencias que para otras: por ejemplo, se podría argumentar que se requiere una cualidad consciente para verdaderamente tener creencias sobre nuestras *experiencias*, y quizá también se requieran ciertos tipos de experiencias para tener ciertos tipos de creencias perceptuales acerca del mundo externo (¿es posible que se necesiten experiencias de color rojo para creer que un objeto es rojo?). En otros casos, esto parece más problemático. Por ejemplo, cuando pienso que Don Bradman es el mejor jugador de criquet de todos los tiempos, es plausible decir que yo habría tenido la misma creencia aun si hubiese tenido una experiencia consciente muy diferente asociada con ello. La fenomenología de la creencia es relativamente tenue, y es difícil ver cómo podría ser esta cualidad fenoménica la que hace de la creencia una creencia sobre Bradman. Lo que parece más importante en el contenido de la creencia es la conexión entre la creencia y Bradman, y el papel que desempeña en mi sistema cognitivo.

Como una posición más débil, podría sugerirse que aunque no se requiere ninguna cualidad fenoménica *particular* para tener una creencia determinada, un ser debe ser *capaz*, al menos, de experiencias conscientes para creer alguna cosa.[11] Existe cierta plausibilidad en la idea de que un ser sin una vida interior consciente no podría verdaderamente ser un creyente; cuanto más sería un pseudocreyente. Aun así, esto haría que el papel de lo fenoménico en los conceptos intencionales sea bastante débil. Los requerimientos más esenciales para tener una creencia específica se encontrarán en otros lados, no en lo fenoménico. Podríamos incluso eliminar cualquier componente fenoménico, dejando un concepto de pseudocreencia que se parece a la creencia en los aspectos más importantes excepto que no involucra el concepto de conciencia. Es plausible que la pseudocreencia pueda realizar gran parte del trabajo explicativo que realiza el concepto de creencia.

En cualquier caso, no intentaré aquí arbitrar estas difíciles cuestiones acerca de la relación entre la intencionalidad y la conciencia. Podemos notar que existe al menos un concepto *deflacionario* de la creencia que es puramente psicológico, y que no involucra a la experiencia consciente; si un ser está en el estado psicológico correcto, entonces está en un estado que se parece a una creencia en muchos modos importantes, excepto en lo que respecta a cualquier aspecto fenoménico. Y existe un concepto *inflacionario* de creencia, en el que se requiere la experiencia consciente para verdaderamente creer, y quizás, incluso, en el que se necesita una clase específica de experiencia consciente para verdaderamente creer una proposición determi-

nada. Cuál de estos es el "verdadero" concepto de creencia no será demasiado importante para mis propósitos.

Lo fundamental es que no hay ninguna característica de las creencias que vaya más allá de lo fenoménico y lo psicológico. Quizá deba hacerse una pequeña salvedad a esta afirmación: podríamos tener que agregar un elemento *relacional*, para dar cuenta del hecho de que ciertas creencias pueden depender del estado del ambiente además del estado interno del pensador. Se argumentó, por ejemplo, que para creer que el agua está mojada, un sujeto debe estar relacionado de un modo apropiado con el agua en el ambiente. Por lo general, se interpreta que esta debe ser una relación causal, de modo que es posible que se pueda incorporar esto en la caracterización de la propiedad psicológica relevante, donde los papeles causales en cuestión se extienden fuera de la cabeza hacia el ambiente. De ser así, no se requeriría entonces ningún componente extra. Pero, en cualquier caso, no es una carga demasiado pesada notar que también podría existir un componente relacional en ciertos estados mentales, además de los componentes psicológicos y fenoménicos. De cualquier manera, no existe ningún profundo misterio ulterior.

Para ver que no existe ningún aspecto profundo ulterior, además de los aspectos fenoménico y psicológico-relacional de los estados intencionales, nótese que los fenómenos manifiestos del mismo tipo que requieren explicación caen en una de dos clases: aquellos a los que tenemos acceso de tercera persona y aquellos a los que tenemos acceso de primera persona. Los que se encuentran en la primera clase se reducen, en última instancia, a conductas, relaciones con el ambiente, etc., y pueden subsumirse en la clase de lo psicológico y lo relacional. Los que se encuentran en la segunda clase se reducen a la *experiencia* asociada con creer —por ejemplo, el modo como nuestros conceptos parecen extenderse a un mundo fenoménico— y así constituye parte del problema de la conciencia, no un misterio separado. Las razones para creer en cualquier aspecto determinado de la creencia (incluyendo el contenido semántico, el "asunto", etc.) surgirán de una de estas dos clases; no existe ninguna tercera clase independiente de fenómenos que estemos obligados a explicar

Otro modo de ver esto es advertir que una vez que hemos fijado las propiedades psicológicas, fenoménicas y relacionales de un individuo, no parece haber ningún otro ente mental que puede variarse en forma independiente. No podemos ni siquiera *imaginar* a alguien que sea idéntico a mí en los tres aspectos mencionados pero que crea algo diferente; sí podemos, en cambio, imaginar a alguien que sea psicológicamente idéntico a mí pero que experimente algo diferente. En el primer caso, simplemente no hay suficiente espacio conceptual

para la posibilidad. Los conceptos intencionales son en cierta forma menos primitivos que los conceptos psicológicos y fenoménicos, en el sentido de que no pueden variarse independientemente de estos últimos.[12]

Todo lo que dije aquí acerca de las creencias se aplica igualmente a otros estados intencionales, como el deseo, la esperanza, etc. Todos ellos tienen un aspecto psicológico y fenoménico, y no necesitamos legislar cuál es el principal, aunque podría hacerse una sólida defensa en favor de un análisis psicológico. Lo importante es que no hay ningún aspecto de este estado que vaya más allá de lo psicológico y lo fenoménico (quizá con un componente relacional añadido). Juntas, la psicología y la fenomenología, constituyen los aspectos centrales de la mente.

## La coocurrencia de las propiedades fenoménicas y psicológicas

Es un hecho acerca de la mente humana que cada vez que se instancia una propiedad fenoménica, se instancia una propiedad psicológica correspondiente. La experiencia consciente no ocurre en el vacío. Está siempre vinculada al procesamiento cognitivo, y es probable que, en algún sentido, surja de ese procesamiento. Cuando tenemos una sensación, por ejemplo, ocurre algún procesamiento de información: la percepción correspondiente, si así lo desea. De modo similar, cada vez que tenemos la experiencia consciente de felicidad, el papel funcional asociado con la felicidad es habitualmente desempeñado por algún estado interno. Quizá sea lógicamente posible que podamos tener experiencias sin causas, pero parece ser un hecho empírico que ocurren juntas.

Ante esta coocurrencia, los temerosos podrían sentirse tentados de poner en duda que estemos haciendo alguna real distinción. Pero es claro que existe por lo menos una distinción conceptual, aunque la extensión de los conceptos involucrados parezca ser la misma. Podemos preguntarnos cómo explicar la cualidad fenoménica, y podemos preguntarnos cómo explicar el desempeño del papel causal, y estas son dos preguntas diferentes.

Dicho esto, la coocurrencia de las propiedades fenoménicas y psicológicas refleja algo profundo acerca de nuestros conceptos fenoménicos. No tenemos ningún lenguaje independiente para describir las cualidades fenoménicas. Como hemos visto, existe algo inefable de ellas. Aunque la verdosidad es un tipo distintivo de sensación con un rico carácter intrínseco, hay muy poco que podamos decir de ella que no sea que es verde. Al hablar acerca de las cualidades

fenoménicas, por lo general debemos especificar las cualidades en cuestión en términos de propiedades externas asociadas, o en términos de papeles causales vinculados. Nuestro *lenguaje* para las cualidades fenoménicas se deriva de nuestro lenguaje no fenoménico. Como dijo Ryle, no hay palabras de sensación "puras".

Si se examina el catálogo de experiencias conscientes que presentamos antes, las experiencias en cuestión nunca se describen en términos de sus cualidades intrínsecas. Más bien, utilicé expresiones como "el aroma del pan recién horneado", "los patrones que se producen cuando cerramos los ojos", etc. Incluso con un término como "sensación de verde", la referencia se fija efectivamente en términos extrínsecos. Cuando aprendemos el término "sensación de verde" es por ostensión: aprendemos a aplicarlo al tipo de experiencia causada por el pasto, los árboles, etc. Por lo general, en la medida en que tenemos categorías fenoménicas comunicables, estas se definen por referencia a sus asociaciones externas típicas o a una clase asociada de estado psicológico. Por ejemplo, cuando hablamos de la cualidad fenoménica de la felicidad, la referencia del término "felicidad" está implícitamente fijada por medio de algún papel causal: el estado por el cual juzgamos que todo está bien, saltos de alegría, etc. Quizás esta sea una interpretación posible de la famosa observación de Wittgenstein, "Un proceso interno tiene necesidad de criterios exteriores".

Esta dependencia de los conceptos fenoménicos sobre criterios causales llevó a algunos autores (incluyendo a Wittgenstein y Ryle, en algunos de sus estados de ánimo) a sugerir que no hay nada en el significado de nuestros conceptos mentales más allá de los criterios causales asociados. Existe una cierta plausibilidad en esto: si siempre se selecciona una propiedad fenoménica invocando una propiedad psicológica, ¿por qué no suponer que sólo hay una propiedad involucrada? Pero debería resistirse esta tentación. Cuando hablamos de una sensación de verde, esta forma de hablar no es equivalente simplemente a hablar de "un estado que es causado por el pasto, los árboles, etc.". Hablamos de la *cualidad fenoménica* que ocurre por lo general cuando un estado es causado por el pasto y los árboles. Si existe un análisis causal cercano, es algo así como "la clase de estado *fenoménico* que es causado por el pasto, los árboles, etc.".[13] El elemento fenoménico en el concepto impide un análisis en términos puramente funcionales.

En general, cuando se selecciona una propiedad fenoménica con la ayuda de una propiedad psicológica $P$, la noción fenoménica no es sólo "$P$". Es "el tipo de experiencia consciente que tiende a acompañar a $P$". Y, lo que es importante, la propia noción de "cualidad fenoménica" o "experiencia consciente" no se define en términos psicológicos. Más

bien, la noción de experiencia consciente es, como vimos antes, una especie de primitivo. *Si* existiese un análisis funcional de la noción de experiencia o cualidad fenoménica, entonces el análisis en cuestión produciría análisis funcionales de propiedades fenoménicas específicas, pero en ausencia de un análisis de este tipo no podemos sacar una conclusión de esta clase.

No podemos identificar la noción "*P* fenoménico" con la de "*P* psicológico" por todas las razones usuales: son dos conceptos distintos, como lo atestigua el hecho de que hay dos explananda distintos. Aunque "*P* fenoménico" se interpreta como "la experiencia que tiende a acompañar a un *P* psicológico", podemos imaginar coherentemente una situación en la cual un *P* fenoménico ocurra sin un *P* psicológico, y viceversa. El icono de un Rolls-Royce puede analizarse aproximadamente como la clase de icono que por lo general se encuentra en los autos Rolls-Royce, pero esto no significa que ser un icono de Rolls-Royce sea ser un auto Rolls-Royce.

Esto nos permite una cierta comprensión acerca de la escasez de nuestro vocabulario específicamente fenoménico en comparación con nuestro vocabulario psicológico, y también nos ayuda a comprender por qué las propiedades fenoménicas y psicológicas han sido fundidas tan frecuentemente. Para la mayoría de los propósitos cotidianos esta fusión no es importante: cuando afirmamos que alguien está alegre, no necesitamos hablar específicamente de la cualidad fenoménica ni del papel funcional, ya que usualmente ambos van juntos. Sin embargo, para los propósitos filosóficos y en particular para los propósitos explicativos, fusionar estas propiedades es fatal. La fusión puede ser tentadora, ya que colapsar la distinción hace que el problema de explicar la experiencia consciente se vuelva repentinamente muy directo; pero, por la misma razón, es por completo insatisfactoria. No podemos hacer desaparecer el problema de la conciencia de un modo puramente verbal.

## 4. Los dos problemas mente-cuerpo

La división de las propiedades mentales en propiedades fenoménicas y psicológicas tiene el efecto de dividir el problema mente-cuerpo en dos: una parte fácil y una parte difícil. Los aspectos psicológicos de la mente plantean muchos problemas técnicos para la ciencia cognitiva, y algunos problemas para el análisis filosófico, pero no plantean ningún profundo enigma metafísico. La pregunta: "¿Cómo podría un sistema físico ser el tipo de ente que puede *aprender* o que puede *recordar*?" no tiene el mismo alcance que la pregunta correspondiente acerca de las sensaciones o de la conciencia en general. La

razón de esto es clara. Según nuestro análisis en el apartado 3, el aprendizaje y la memoria son propiedades funcionales caracterizadas por papeles causales, de modo que la pregunta "¿Cómo puede un sistema físico tener una propiedad psicológica P?" se reduce a lo mismo que "¿Cómo puede un estado de un sistema físico desempeñar tal y cual papel causal?" Esta es una pregunta para las ciencias de los sistemas físicos. Simplemente debemos contar una historia sobre cómo la organización del sistema físico le permite reaccionar a la estimulación ambiental y producir conductas del tipo apropiado. Aunque los problemas técnicos son enormes, existe un programa de investigación claramente definido para dar respuesta a esa pregunta. Los problemas metafísicos son relativamente pocos.

Esto no significa que las propiedades psicológicas *no* planteen dificultades filosóficas. Por ejemplo, existen problemas significativos para lograr el análisis correcto de estas nociones. Aunque por lo general se acepta que se trata de conceptos funcionales, pueden existir importantes discrepancias sobre cómo deberían hacerse los análisis funcionales requeridos. Las propiedades intencionales como la creencia y el deseo, por ejemplo, son un terreno fértil para la discusión. En particular, la cuestión de qué constituye el contenido de un estado intencional determinado todavía se comprende de un modo insuficiente. Existen también problemas técnicos con respecto a cómo pueden las construcciones de alto nivel como estas desempeñar un papel causal real en la producción de la conducta, especialmente cuando están constituidas en parte por propiedades del ambiente, o acerca de la existencia de leyes estrictas que conecten los estados psicológicos con la conducta. Existen, además, problemas semiempíricos en la fundamentación de la ciencia cognitiva vinculados a exactamente cómo deberían instanciarse estas propiedades en los sistemas cognitivos existentes, o incluso a la posibilidad de su instanciación.

Estos interrogantes son todos importantes, pero tienen el carácter de problemas más que de misterios. La situación aquí es análoga a la de la filosofía de la biología, para la cual no existe ningún problema vida-cuerpo acuciante; hay meramente una multitud de problemas técnicos sobre la evolución, la selección, la adaptación, la aptitud, y las especies. Así como la mayoría de los aparentes misterios metafísicos que rodeaban a la biología fueron eliminados hace mucho, es justo decir que el problema mente-cuerpo para las propiedades psicológicas se ha diluido para todo propósito. Lo que resta es una colección de problemas técnicos más pequeños que el curso normal del análisis científico y filosófico puede manejar.

Los aspectos fenoménicos de la mente son una cuestión diferente. Aquí el problema mente-cuerpo es tan desconcertante como

siempre. El impresionante progreso de las ciencias físicas y cognitivas no arrojó ninguna luz significativa sobre la cuestión de cómo y por qué el funcionamiento cognitivo está acompañado por la experiencia consciente. El progreso en la comprensión de la mente se concentró casi exclusivamente en la explicación de la conducta. Este progreso no incluye la cuestión de la experiencia consciente.

Si así nos place, podemos considerar la distinción psicológico-fenoménica no tanto como dividir el problema mente-cuerpo sino como factorizarlo en dos partes separadas. La parte más difícil del problema mente-cuerpo se encuentra en la pregunta: ¿cómo podría un sistema físico originar la experiencia consciente? Podríamos factorizar el vínculo entre la experiencia física y consciente en dos partes: el vínculo entre lo físico y lo psicológico, y el vínculo entre lo psicológico y lo fenoménico. Como vimos más arriba, en este momento tenemos una idea bastante buena de cómo un sistema físico puede tener propiedades psicológicas: el problema *psicológico* mente-cuerpo se ha disuelto. Lo que persiste es la cuestión de por qué y cómo estas propiedades psicológicas están acompañadas por propiedades fenoménicas: por qué, por ejemplo, toda estimulación y reacción asociada con el dolor está acompañada por la *experiencia* del dolor. Siguiendo a Jackendoff (1987), podemos llamar a este residuo el *problema mente-mente*. Las explicaciones físicas actuales nos llevan hasta la mente psicológica. Lo que permanece mal comprendido es el vínculo entre la mente psicológica y la mente fenoménica.[14]

Es concebible que el vínculo entre lo fenoménico y lo físico pudiese ser independiente del vínculo entre lo psicológico y lo físico, de modo que la factorización fuera imposible, pero esto parece improbable. La estrecha correlación que hemos visto entre las propiedades fenoménicas y psicológicas sugiere un vínculo profundo. En capítulos posteriores argumentaré que el vínculo es extremadamente fuerte y que la estrategia de factorización es valiosa para aproximarse al problema mente-cuerpo. Si esto es así, entonces la comprensión del vínculo entre lo psicológico y lo fenoménico es crucial para comprender la experiencia consciente.

## 5. Dos conceptos de conciencia

Si se tiene en cuenta que tantos términos mentales poseen una naturaleza dual, no resultará sorprendente descubrir que incluso la "conciencia" tiene un sentido fenoménico y otro psicológico. Hasta ahora nos hemos concentrado en el sentido fenoménico, el que a su vez subsume a todos los aspectos fenoménicos previamente mencionados de la mente. Ser consciente en este sentido es sólo instanciar alguna

cualidad fenoménica. Este es el sentido clave de "conciencia" o, al menos, el que plantea los principales problemas explicativos. Pero no es el único sentido del término. "Conciencia" puede usarse también para referir a una variedad de propiedades psicológicas, como la informatividad o la accesibilidad introspectiva a la información. Podemos agrupar las propiedades psicológicas de este tipo bajo el rótulo de *conciencia psicológica*, para distinguirla de la *conciencia fenoménica* de la que me he estado ocupando.

Esta ambigüedad puede llevar a mucha confusión cuando se discute sobre la conciencia. Es frecuente que alguien que formula una explicación de la conciencia comience revistiendo al problema de toda la gravedad del problema de la conciencia fenoménica, pero termine dando una explicación de algún aspecto de la conciencia psicológica, tal como la capacidad de hacer introspección. Esta explicación puede ser valiosa por derecho propio, pero nos queda la sensación de que se nos prometió más de lo que se cumplió.

## Variedades de la conciencia psicológica

Existen numerosas nociones psicológicas para las que se utiliza a veces el término "conciencia". Estas incluyen las siguientes:

*Vigilia*. A veces decimos que una persona está consciente como una forma de decir que no está dormida. Tiene sentido suponer que tenemos experiencias mientras estamos dormidos, de modo que esta noción claramente no coincide con la conciencia fenoménica. Es plausible analizar la vigilia en términos funcionales, quizás, en una primera aproximación, en términos de una capacidad para procesar información acerca del mundo y tratar con ella de un modo racional.

*Introspección*. Este es el proceso por el cual podemos volvernos conscientes del contenido de nuestros estados internos. Si usted me pregunta acerca de mis estados mentales, es por introspección que yo determino mi respuesta. Este acceso a los propios estados mentales es un componente importante del concepto cotidiano de conciencia y es, al menos parcialmente, una noción funcional. Podríamos analizarlo en términos de que nuestros procesos racionales son sensibles de la forma correcta a la información acerca de nuestros propios estados internos, y que podemos utilizar esa información en forma apropiada.

*Informatividad*. Esta es nuestra capacidad para informar sobre el contenido de nuestros estados mentales. Presupone la capacidad de hacer introspección, pero está más constreñida que esa capacidad, ya que reconoce la posibilidad de utilizar el lenguaje. Este concepto de

conciencia ha sido frecuentemente el blanco central de los filósofos y psicólogos de tendencia operacionalista.

*Autoconciencia*. Esta refiere a nuestra capacidad para pensar sobre nosotros mismos, la conciencia de nuestra existencia como individuos y de nuestras diferencias de los otros. Mi autoconciencia podría analizarse en términos de mi acceso a un modelo propio, o mi posesión de un cierto tipo de representación que está asociada de algún modo conmigo mismo. Bien podría ocurrir que la autoconciencia esté limitada a los seres humanos y a unas pocas especies animales.

*Atención*. Con frecuencia decimos que alguien es consciente de algo precisamente cuando presta atención a ello; esto es, cuando una porción significativa de sus recursos cognitivos está dedicada a tratar con la información relevante. Podemos ser fenoménicamente conscientes de algo sin atender a ello, como lo atestigua la periferia del campo visual.

*Control voluntario*. En otro sentido, decimos que un *acto* conductual es consciente cuando lo realizamos deliberadamente; esto es, cuando la acción es causada del modo apropiado por un elemento de pensamiento previo.

*Conocimiento*. En otro sentido cotidiano, decimos que alguien es consciente de un hecho precisamente cuando conoce el hecho, y que es consciente de una cosa precisamente cuando conoce acerca de esa cosa. Esta noción es raramente el centro de la discusión técnica de la conciencia, pero probablemente es tan fundamental para el uso cotidiano del término como cualquier otra.

Todas estas son nociones principalmente funcionales. Esto puede verse a través de cómo explicaríamos los fenómenos en cuestión. Si intentásemos explicar la atención, podríamos diseñar un modelo de los procesos cognitivos que llevan a concentrar recursos en un aspecto de la información disponible en lugar de en otro. Si fuésemos a intentar explicar la introspección, trataríamos de aclarar los procesos mediante los cuales somos sensibles a nuestros estados internos del modo apropiado. Historias similares se aplican a las explicaciones de las otras propiedades. En cada caso, la explicación funcional parece capturar lo que es fundamental.

Aunque estos conceptos tienen un núcleo psicológico, muchos o todos ellos están asociados a estados fenoménicos. Por ejemplo, existe un cierto tipo de estado fenoménico asociado con la autoconciencia. Lo mismo ocurre para la introspección, la atención y el control voluntario de la conciencia. Como con los otros términos de aspecto dual que ya hemos discutido, los términos como "introspección" y "autoconciencia" también se utilizan a veces para referir al estado fenoménico, lo que

puede prestarse a confusión. Alguien podría argumentar que se requiere un aspecto fenoménico para que un proceso califique verdaderamente de “introspección”, “atención”, o lo que sea. Como antes, sin embargo, esta cuestión es predominantemente verbal. Es claro que existe una propiedad fenoménica y una psicológica en la vecindad de cada uno de estos conceptos. Aquellos a los que no les gusta dignificar la propiedad psicológica con un término mental como “atención” pueden utilizar la palabra “pseudoatención” en cambio. Las cuestiones filosóficas substanciales siguen siendo las mismas, independientemente de cómo se denominen las propiedades.

Las propiedades fenoménicas y psicológicas cercanas a estas nociones tienden a ocurrir juntas, pero como con otros conceptos mentales, no deberían fusionarse. También debemos ser cuidadosos de no fundir los sentidos fenoménicos de estos términos con la conciencia fenoménica en general.

## Conciencia y percatación

Hemos visto que existe una propiedad psicológica asociada con la experiencia de emoción, una propiedad psicológica asociada con la experiencia de autoconciencia, una propiedad psicológica asociada con la experiencia de sensación, etc. Es natural suponer que podría haber una propiedad psicológica asociada con la propia experiencia, o con la conciencia fenoménica. De hecho, creo que existe cerca una propiedad semejante; podemos llamarla “percatación”. Esta es la marca más general de la conciencia psicológica.

La *percatación* puede analizarse en un sentido amplio como un estado en el cual tenemos acceso a alguna información, y podemos usar esa información en el control de la conducta. Podemos percatarnos, entre otras cosas, de un objeto en el ambiente, de un estado de nuestro propio cuerpo o de un estado mental. La percatación de la información por lo general conlleva la capacidad de, a sabiendas, dirigir la conducta según esa información. Esta es claramente una noción funcional. En el lenguaje cotidiano, el término “percatación” se usa frecuentemente como sinónimo de “conciencia”, pero reservaré el término para la noción funcional que he descripto aquí.

En general, dondequiera que exista conciencia fenoménica parece haber percatación. Mi experiencia fenoménica del libro amarillo al lado mío está acompañada por mi percatación funcional del libro, y ciertamente por mi percatación del color amarillo. Mi experiencia de un dolor está acompañada por una percatación de la presencia de algo desagradable que tiende a conducir al alejamiento o algo parecido, cuando ello es posible. El hecho de que cualquier experiencia cons-

ciente esté acompañada por la percatación se hace evidente por el hecho de que una experiencia consciente es *comunicable*. Si estoy teniendo una experiencia, puedo hablar acerca de que la estoy teniendo. Puedo no estar prestando atención a ella, pero al menos tengo la capacidad de concentrarme en ella y hablar de ella, si así lo deseo. Esta informatividad implica inmediatamente que me percato de ella en el sentido relevante. Por supuesto, un animal o ser humano prelingüístico podría tener una experiencia consciente sin la capacidad para informarla, pero es plausible que esta clase de ser posea de todos modos un cierto grado de percatación. La percatación no supone la capacidad de informar, aunque las dos tienden a ocurrir conjuntamente en las criaturas con lenguaje.

La conciencia está siempre acompañada de percatación, pero la percatación tal como la he descripto no necesita estar acompañada de conciencia. Por ejemplo, podemos percatarnos de un hecho sin ninguna experiencia fenoménica particular asociada. Sin embargo, podría ser posible constreñir la noción de percatación de modo que resulte coextensional, o casi, con la conciencia fenoménica. No intentaré realizar aquí este proyecto, pero lo retomaré en el capítulo 6.

La noción de percatación subsume a la mayoría o a todas las diversas nociones psicológicas de conciencia recién enumeradas. La introspección puede analizarse como percatación de algún estado interno. La atención, como un grado de percatación particularmente alto de un objeto o suceso. La autoconciencia se entendería como la percatación de uno mismo. El control voluntario es más complicado, pero podría analizarse en parte como un requisito de atención a la conducta que estamos realizando. Sería posible caracterizar aproximadamente a la vigilia como un estado en el cual podemos tratar de un modo racional con nuestro ambiente, por lo que implica alguna clase de percatación.

La idea de que existe una noción funcional de conciencia que puede explicarse en términos de acceso fue elaborada por Block (1995), quien habla de la distinción entre la "conciencia fenoménica" y la "conciencia de acceso". La noción de Block de conciencia de acceso corresponde estrechamente a la noción de percatación que he estado describiendo (analizaré la relación más en profundidad en el capítulo 6). De un modo similar, Newell (1992) distingue explícitamente entre "percatación" y "conciencia". Describe la percatación como "la capacidad de un sujeto para hacer que su conducta dependa de algún conocimiento" y elabora la distinción entre esa noción y la conciencia, que, según él, es un fenómeno no funcional. Otros filósofos y científicos cognitivos hicieron distinciones similares.[15]

La percatación, como otras tantas propiedades psicológicas, plantea pocos problemas metafísicos. Los problemas planteados por las variedades psicológicas de la conciencia son del mismo orden de magnitud que aquellos planteados por la memoria, el aprendizaje y la creencia. Ciertamente, la noción de percatación no es cristalina, de modo que hay espacio para análisis filosóficos significativos sobre lo que realmente quiere decir. Más aún, hay espacio para una enorme cantidad de investigaciones en ciencia cognitiva sobre cómo los sistemas cognitivos naturales y artificiales podrían funcionar para poseer la capacidad de percatación. Pero las líneas generales de estos programas de investigación son razonablemente claras. Hay pocas razones para suponer que el curso normal de la ciencia cognitiva, respaldada por un análisis filosófico apropiado, no alcance el éxito con el tiempo.

En la medida en que la conciencia es el problema realmente difícil de una ciencia de la mente, la conciencia fenoménica ocupará un lugar central. Los problemas aquí son de un orden de magnitud diferente. Aun después de haber explicado el funcionamiento físico y computacional de un sistema consciente, deberemos explicar por qué el sistema tiene experiencias conscientes. Por supuesto, algunos cuestionan esta afirmación, razón por lo cual en breve la justificaré más ampliamente. Por ahora, sin embargo, simplemente hacemos notar la diferencia *prima facie* en los problemas que las variedades fenoménica y psicológica presentan. Es la conciencia fenoménica la que plantea el problema *preocupante* de la conciencia.

Considerando las diferencias entre las nociones psicológica y fenoménica de la conciencia, resulta desafortunado que frecuentemente se las fusione en la literatura. Esta fusión importa poco en el habla cotidiana, ya que la percatación y la conciencia fenoménica por lo general ocurren juntas. Pero, para los propósitos de la explicación, es crucial su distinción conceptual. En la medida en que se hayan formulado explicaciones aunque más no sea remotamente satisfactorias de la "conciencia", lo que se explica es, por lo general, un aspecto psicológico. Usualmente no se consideran los aspectos fenoménicos.

Son muchos los análisis filosóficos recientes de la conciencia que se ocuparon principalmente de los aspectos no fenoménicos. Rosenthal (1996) argumenta que un estado mental es consciente precisamente cuando existe un pensamiento de orden superior sobre ese estado mental. Este podría ser un análisis útil de la conciencia introspectiva y, quizá, de otros aspectos de la percatación, pero no parece explicar la experiencia fenoménica.[16] De modo similar, Dennett (1991) utiliza

buena parte del libro para delinear un modelo cognitivo detallado, que formula como una explicación de la conciencia. A primera vista, el modelo es fundamentalmente un modelo de la capacidad de un sujeto para informar en forma verbal sobre un estado mental. De este modo, podría dar una explicación de la informatividad, de la conciencia introspectiva, y quizá de otros aspectos de la percatación, pero nada en el modelo proporciona una explicación de la conciencia fenoménica (aunque Dennett pondría las cosas de otra forma).

Armstrong (1968), para quien la conciencia es un obstáculo para su teoría funcionalista de la mente, analiza la noción en términos de la presencia de algún mecanismo de autoexploración. Esto podría proporcionar una concepción útil de la autoconciencia y de la conciencia introspectiva, pero deja de lado el problema de la experiencia fenoménica. Armstrong (1981) habla sobre la conciencia perceptual y la conciencia introspectiva, pero sólo se ocupa de ellas en tanto variedades de la percatación, y no enfrenta los problemas planteados por las cualidades fenoménicas de la experiencia. De esta forma, y gracias a la ambigüedad de la noción de conciencia, deja de lado el sentido de conciencia realmente problemático para su teoría funcionalista.

Otros autores que escribieron sobre el tema de la "conciencia" se ocuparon principalmente de la autoconciencia o de la conciencia introspectiva. Van Gulick (1988), al sugerir que la conciencia debería analizarse como la posesión de "información metapsicológica reflexiva", proporciona cuanto más un análisis de estas nociones psicológicas y ciertamente acepta que los aspectos fenoménicos pueden ser dejados de lado por un análisis de este tipo. De forma similar, la elaborada teoría de Jaynes (1976) sobre la conciencia sólo se ocupa de la percatación de nuestros propios pensamientos. No dice nada sobre los fenómenos asociados con la percepción y por lo tanto no puede pretender ser una teoría de la percatación en general, y mucho menos una teoría de la conciencia fenoménica. Hofstadter (1979) tiene algunas cosas interesantes para decir acerca de la conciencia, pero se ocupa más de la introspección, el libre albedrío y el sentido del sí mismo que de la experiencia *per se*.

Si bien la conciencia ha sido un tema de discusión entre los psicólogos, las nociones fenoménica y psicológica por lo general no han sido distinguidas cuidadosamente. Por lo general es algún aspecto de la percatación, como la introspección, la atención o la autoconciencia, lo que esos estudios psicológicos encaran. Incluso, los aspectos psicológicos de la conciencia han tenido algo así como una mala reputación en psicología, al menos hasta hace poco. Quizás, esto se deba a cierta confusión sobre esas nociones, y a las dificultades

asociadas con fenómenos de alto nivel como la introspección. Se podría especular que esa mala reputación se debe, en buena medida, a que comparte el nombre con la conciencia fenoménica, dando la apariencia de complicidad en un crimen.

A veces se escucha que la investigación psicológica ha "vuelto a la conciencia" en años recientes. La realidad parece ser que los aspectos psicológicos de la conciencia han sido un tema activo de investigación, y que los investigadores no han tenido miedo de utilizar el término "conciencia" para el fenómeno. Por lo general, sin embargo, se sigue dejando de lado a la conciencia fenoménica. Esto es comprensible. Sabemos cómo los métodos de la psicología experimental podrían llevarnos a una comprensión de los diversos tipos de percatación, pero no nos resulta fácil concebir cómo estos podrían explicar la experiencia fenoménica.[17]

Los modelos cognitivos están bien preparados para explicar los aspectos psicológicos de la conciencia. No existe ningún vasto problema metafísico en la idea de que un sistema físico debería poder hacer introspección de sus estados internos, o manejar racionalmente la información de su ambiente, o ser capaz de concentrar su atención primero en un sitio y luego en el próximo. Es suficientemente claro que una concepción funcional apropiada debería poder explicar estas capacidades, aun cuando el descubrimiento de la teoría correcta podría demorar décadas o siglos. Pero el problema realmente difícil es el de la conciencia fenoménica, y este no ha sido tocado por las explicaciones de la conciencia psicológica formuladas hasta ahora.

En lo que sigue, utilizaré el término "conciencia" para referir sólo a la conciencia fenoménica. Cuando desee utilizar las nociones psicológicas, hablaré de "conciencia psicológica" o "percatación". Es de la conciencia fenoménica de lo que me ocuparé principalmente.

# 2

# Superveniencia y explicación

¿Cuál es el lugar de la conciencia en el orden natural? ¿La conciencia es física? ¿Puede la conciencia explicarse en términos físicos? Para poder responder estas preguntas, necesitamos construir un marco teórico; eso es lo que haremos en este capítulo. La pieza central de este marco teórico es el concepto de *superveniencia*: formularé una definición de este concepto y la utilizaré para aclarar la idea de explicación reductiva. Utilizando esta concepción, trazaré una imagen de la relación entre la mayoría de los fenómenos de alto nivel y los hechos físicos, una que parece cubrir todo excepto, quizá, la experiencia consciente.

## 1. Superveniencia

Se encuentra muy difundida la creencia de que los hechos más básicos acerca de nuestro universo son hechos físicos y que todos los demás hechos dependen de ellos. En un sentido bastante débil de "depender" esto podría ser casi trivialmente verdadero; en un sentido fuerte de "depender", esto es controversial. Existe una compleja variedad de relaciones de dependencia entre los hechos de alto nivel y los hechos de bajo nivel en general, y el tipo de relación de dependencia válido en un dominio, como la biología, puede no ser válido en otro, como el de la experiencia consciente. La noción filosófica de superveniencia proporciona un marco unificador dentro del cual pueden analizarse estas relaciones de dependencia.

La noción de superveniencia formaliza la idea intuitiva de que un conjunto de hechos puede determinar por completo otro conjunto de hechos.[1] Por ejemplo, los hechos físicos acerca del mundo parecen determinar a los hechos biológicos, en el sentido de que una vez que todos los hechos físicos fueron fijados, no hay lugar para que los

hechos biológicos varíen. (Fijar todos los hechos físicos simultáneamente determinará qué objetos están vivos.) Esto proporciona una caracterización aproximada del sentido en el cual las propiedades biológicas supervienen a las propiedades físicas. En general, la superveniencia es una relación entre dos conjuntos de propiedades: propiedades *B* —intuitivamente, las propiedades de *alto nivel*— y propiedades *A*, que son las propiedades más básicas de *bajo nivel*.

Para nuestros propósitos, las propiedades A relevantes son por lo general las propiedades físicas: más precisamente, las propiedades fundamentales invocadas por una teoría terminada de la física. Quizás estas incluyan la masa, la carga, la posición espaciotemporal; las propiedades que caracterizan la distribución de diversos campos espaciotemporales, la aplicación de diversas fuerzas, y la forma de diversas ondas; etc. La naturaleza precisa de estas propiedades no es importante. Si la física cambiase radicalmente, la clase relevante de propiedades podría ser bastante diferente de aquellas que menciono, pero los argumentos seguirán siendo válidos de todas maneras. Propiedades de alto nivel como la jugosidad, la grumosidad, la jiraficidad y otras están excluidas, aunque existe un sentido en el cual estas propiedades son físicas. En lo que sigue, nuestro discurso sobre las propiedades físicas estará implícitamente restringido a la clase de propiedades fundamentales, a menos que se indique lo contrario. A veces hablaré de las propiedades "microfísicas" o "físicas de bajo nivel" para ser explícito.

Los *hechos A* y *hechos B* acerca del mundo son los hechos concernientes a la instanciación y distribución de propiedades A y propiedades B.[2] De modo que los hechos físicos acerca del mundo abarcan todos los hechos relativos a la instanciación de propiedades físicas dentro del continuo espaciotemporal. También es útil estipular que los hechos físicos del mundo incluyen sus leyes físicas básicas. En algunas concepciones, estas leyes ya están determinadas por la totalidad de los hechos físicos particulares, pero no podemos darlo por sentado.

La plantilla para la definición de superveniencia es la siguiente:

> Las propiedades B *supervienen* a las propiedades A si ningún par de situaciones posibles es idéntico respecto de sus propiedades A pero difiere en sus propiedades B.

Por ejemplo, las propiedades biológicas supervienen a las propiedades físicas si dos situaciones cualesquiera posibles que son físicamente idénticas son también biológicamente idénticas. (Aquí utilizo "idéntico" en el sentido de indiscernibilidad más que en el de

identidad numérica. En este sentido, dos mesas distintas podrían ser físicamente idénticas.) Pueden obtenerse nociones más precisas de superveniencia llenando esta plantilla. Según interpretemos las "situaciones" en cuestión como individuos o mundos enteros, obtendremos las nociones de superveniencia *local* o *global*, respectivamente. Y según cómo interpretemos la noción de posibilidad, obtendremos las nociones de superveniencia *lógica*, superveniencia *natural* y quizás otras. Especificaré estas distinciones en lo que sigue.

## Superveniencia local y global

Las propiedades B supervienen *localmente* a las propiedades A si las propiedades A de un *individuo* determinan las propiedades B de ese mismo individuo, esto es, si dos individuos cualesquiera posibles que instancian las mismas propiedades A instancian también las mismas propiedades B. Por ejemplo, la forma superviene localmente a las propiedades físicas: dos objetos cualesquiera con las mismas propiedades físicas tendrán necesariamente la misma forma. El valor, sin embargo, no superviene localmente a las propiedades físicas: una réplica física exacta de la Mona Lisa no tiene el mismo valor que esta. En general, no existe superveniencia local a lo físico de una propiedad si esa propiedad de alguna forma depende del contexto, esto es, si la posesión de esa propiedad por parte del objeto depende no sólo de su constitución física sino también de su ambiente y su historia. La Mona Lisa es más valiosa que su réplica debido a una diferencia en su contexto histórico: la Mona Lisa fue pintada por Leonardo, pero la réplica no.[3]

En cambio, las propiedades B supervienen *globalmente* a las propiedades A si los hechos A acerca de todo el *mundo* determinan los hechos B: esto es, si no hay dos mundos posibles que son idénticos respecto de sus propiedades A, pero que difieren respecto de sus propiedades B.[4] Un mundo aquí debe pensarse como todo un universo; diferentes mundos posibles corresponden a diferentes modos en los que el universo podría ser.

La superveniencia local implica la superveniencia global, pero no viceversa. Por ejemplo, es plausible que las propiedades biológicas supervengan globalmente a las propiedades físicas, en el sentido de que cualquier mundo físicamente idéntico al nuestro también será biológicamente idéntico. (Sin embargo, es necesario hacer aquí alguna pequeña advertencia, que expondré en unos momentos.) Pero probablemente no supervienen localmente. Es posible que dos organismos físicamente idénticos difieran en ciertas características biológicas. Por ejemplo, uno podría ser más *apto* que el otro debido a

diferencias en sus contextos ambientales. Es concebible, incluso, que organismos físicamente idénticos pudieran ser miembros de diferentes especies si hubiesen tenido historias evolutivas diferentes.

La distinción entre superveniencia global y local no es demasiado importante cuando se trata de la experiencia consciente, porque es probable que si la conciencia superviene a lo físico, lo haga en forma local. Si dos criaturas son físicamente idénticas, entonces las diferencias en los contextos ambientales e históricos no les impedirán tener experiencias idénticas. Por supuesto, el contexto puede afectar la experiencia en forma indirecta, pero sólo en virtud de que afecta la estructura interna, como en el caso de la percepción. Los fenómenos como las alucinaciones y las ilusiones ilustran el hecho de que es su estructura interna y no el contexto lo que es directamente responsable de la experiencia.

### Superveniencia lógica y natural

Una distinción más importante para nuestros propósitos es la que existe entre la superveniencia *lógica* (o conceptual) y la mera superveniencia *natural* (o nomológica, o empírica).

Las propiedades B supervienen *lógicamente* a las propiedades A si ningún par de situaciones *lógicamente posibles* son idénticas respecto de sus propiedades A pero distintas respecto de sus propiedades B. Tendré más para decir acerca de la posibilidad lógica más adelante en este capítulo. Por ahora, se la puede considerar como posibilidad en el sentido más amplio de la palabra, lo que corresponde aproximadamente a la conceptibilidad, relativamente no constreñida por las leyes de nuestro mundo. Es útil pensar en un mundo lógicamente posible como un mundo que podría haber estado dentro de las posibilidades de creación de Dios (¡hipotéticamente!), si así lo hubiese deseado.[5] Dios no podría haber creado un mundo con zorras machos, pero podría haber creado un mundo con teléfonos voladores. En la determinación de la posibilidad lógica de que algún enunciado sea verdadero, las restricciones son principalmente *conceptuales*. La noción de una zorra macho es contradictoria, de modo que es una noción lógicamente imposible; la noción de un teléfono volador es conceptualmente coherente, aunque un poco fuera de lo ordinario, de modo que un teléfono volador es lógicamente posible.

Debería ponerse de relieve que la superveniencia lógica no se define en términos de deducibilidad en cualquier sistema de lógica formal. Más bien, la superveniencia lógica se define en términos de *mundos* (e individuos) lógicamente posibles, donde la noción de un mundo lógicamente posible es independiente de estas consideraciones formales. Este tipo de posibilidad suele denominarse en la

literatura filosófica posibilidad "lógica en sentido amplio", en oposición a la posibilidad "lógica en sentido estricto" que depende de sistemas formales.[6]

En el nivel global, las propiedades biológicas supervienen lógicamente a las propiedades físicas. Ni siquiera Dios podría haber creado un mundo que fuese físicamente idéntico al nuestro pero biológicamente distinto. Simplemente no hay espacio lógico para que los hechos biológicos varíen en forma independiente. Cuando fijamos todos los hechos físicos acerca del mundo —incluyendo los hechos sobre la distribución de cada partícula a través del espacio y tiempo— habremos también fijado la forma macroscópica de todos los objetos en el mundo, el modo como se mueven y funcionan, el modo como interactúan físicamente. Si hay un canguro viviente en este mundo, entonces cualquier mundo que sea físicamente idéntico a este contendrá un canguro físicamente idéntico y ese canguro automáticamente estará vivo.

Podemos imaginar que un superser hipotético —digamos el demonio de Laplace, que conoce la posición de cada partícula en el universo— podría directamente "leer" todos los hechos biológicos, una vez que conoce todos los hechos microfísicos. Estos son suficientes para que este ser construya un modelo de la estructura microscópica y la dinámica del mundo todo a lo largo del espacio y tiempo, a partir de lo cual puede directamente deducir la estructura macroscópica y su dinámica. Con esto, tiene toda la información que necesita para determinar qué sistemas están vivos, qué sistemas pertenecen a la misma especie, etc. Si posee los conceptos biológicos y una especificación completa de los hechos microfísicos, ninguna otra información será relevante.

En general, cuando las propiedades B supervienen lógicamente a las propiedades A, podemos decir que los hechos A *implican* a los hechos B, donde un hecho implica a otro si es lógicamente imposible que el primero sea verdadero y el segundo no. En estos casos, el demonio de Laplace podría leer los hechos B a partir de una especificación de los hechos A, siempre que posea los conceptos B en cuestión. (Diré mucho más acerca de las conexiones entre estos diferentes modos de entender la superveniencia lógica más adelante en este mismo capítulo; el presente análisis es fundamentalmente con propósitos de ilustración.) En cierto sentido, cuando la superveniencia lógica es válida, *todo lo que implica* que los hechos B sean como son es que los hechos A son como son.

Puede haber superveniencia sin superveniencia lógica, sin embargo. La variedad más débil de superveniencia surge cuando dos conjuntos de propiedades están sistemática y perfectamente

*correlacionados* en el mundo natural. Por ejemplo, la presión que ejerce un mol de gas depende sistemáticamente de su temperatura y volumen según la ley $pV = KT$, donde K es una constante (para los propósitos de la ilustración supongo que todos los gases son ideales). En el mundo real, cuando hay un mol de gas a una temperatura y volumen dados, su presión estará determinada: es empíricamente imposible que dos moles distintos de gas puedan tener la misma temperatura y volumen pero diferente presión. Se deduce que, en cierto sentido, la presión de un mol de gas superviene a su temperatura y volumen. (En este ejemplo, interpreto a la clase de propiedades A en una forma mucho más estrecha que a la clase de propiedades físicas, por razones que se aclararán pronto.) Pero esta superveniencia es más débil que la superveniencia lógica. Es *lógicamente* posible que un mol de gas con una temperatura y volumen determinados puedan tener una presión diferente; por ejemplo, imagínese un mundo en el cual la constante K del gas sea mayor o menor. Más bien, es sólo un hecho de la *naturaleza* que exista esa correlación.

Este es un ejemplo de superveniencia *natural* de una propiedad a otras: en esta instancia, la presión superviene naturalmente a la temperatura, al volumen y a la propiedad de ser un mol de gas. En general, las propiedades B supervienen naturalmente a las propiedades A si cualesquiera dos situaciones *naturalmente posibles* con las mismas propiedades A tienen las mismas propiedades B.

Una situación naturalmente posible es una que podría realmente ocurrir en la naturaleza sin violar ninguna ley natural. Esta es una restricción mucho más fuerte que la mera posibilidad lógica. Por ejemplo, un escenario con una constante diferente para el gas es lógicamente posible, pero nunca podría ocurrir en el mundo real, de modo que no es naturalmente posible. Entre las situaciones naturalmente posibles, dos moles de gas cualquiera con la misma temperatura y volumen tendrán la misma presión.

De modo intuitivo, la posibilidad natural corresponde a lo que consideramos es una posibilidad *empírica* real: una situación naturalmente posible es una que podría surgir en el mundo real, si las condiciones fuesen las correctas. Estas incluyen no sólo situaciones reales sino también situaciones contrafácticas que podrían haber surgido en la historia del mundo, si las condiciones fronterizas hubiesen sido diferentes, o que podrían surgir en el futuro, según cómo resulten las cosas. Por ejemplo, un rascacielos de dos kilómetros de alto casi seguro es naturalmente posible, aun cuando en la actualidad no se haya construido ninguno. Incluso es naturalmente posible (aunque muy improbable) que un mono pudiese mecanografiar *Hamlet*. También podemos considerar que una situación naturalmente posible

es una situación conforme a las leyes naturales de nuestro mundo.[7] Por esta razón, la posibilidad natural se denomina a veces posibilidad *nomológica*,[8] proveniente del término griego *nomos*, "ley".

Existe un vasto número de situaciones lógicamente posibles que no son naturalmente posibles. Cualquier situación que viola las leyes naturales de nuestro mundo cae dentro de esta clase: un universo sin gravedad, por ejemplo, o con valores diferentes de las constantes fundamentales. La ciencia ficción proporciona muchas situaciones de este tipo, tales como dispositivos antigravitacionales y las máquinas de movimiento perpetuo. Estas son fáciles de imaginar, pero casi seguramente nunca existirán en nuestro mundo.

En la dirección opuesta, cualquier situación que sea naturalmente posible es también lógicamente posible. La clase de las posibilidades naturales es, por lo tanto un subconjunto de la clase de posibilidades lógicas. Para ilustrar esta distinción: un kilómetro cúbico de oro y un kilómetro cúbico de uranio 235 parecen lógicamente posibles pero, por lo que sabemos, sólo el primero es naturalmente posible, un kilómetro cúbico (estable) de uranio 235 no podría existir en nuestro mundo.

La superveniencia natural ocurre cuando, entre todas las situaciones naturalmente posibles, las que poseen la misma distribución de propiedades A tienen la misma distribución de propiedades B: esto es, cuando los hechos A acerca de una situación *naturalmente necesitan* los hechos B. Esto ocurre cuando las mismas agrupaciones de propiedades A en nuestro mundo están siempre acompañadas por las propiedades B, y cuando esta correlación no es sólo accidental sino *legaliforme*: esto es, la instanciación de las propiedades A siempre producirá las propiedades B, cuando y donde sea que esto ocurra. (En términos filosóficos, la dependencia debe mantenerse en los contrafácticos.) Esta coocurrencia no tiene por qué mantenerse en toda situación lógicamente posible, pero debe ser válida en toda situación naturalmente posible.

Es claro que la superveniencia lógica implica la superveniencia natural. Si dos situaciones lógicamente posibles cualesquiera con las mismas propiedades A tienen las mismas propiedades B, entonces dos situaciones cualesquiera naturalmente posibles también lo harán. Lo recíproco, sin embargo, no es válido, como lo ilustra la ley de los gases. La temperatura y el volumen de un mol de gas determinan su presión en todas las situaciones naturalmente posibles, pero no en todas las lógicamente posibles, de modo que la presión depende naturalmente pero no lógicamente de la temperatura y el volumen. Cuando tenemos superveniencia natural sin superveniencia lógica, diremos que tenemos una *mera* superveniencia natural.

Por razones que se aclararán más adelante, es difícil encontrar casos de superveniencia natural en el conjunto de propiedades *físicas* sin superveniencia lógica, pero la conciencia puede proporcionar una ilustración útil. Parece muy probable que la conciencia sea naturalmente superveniente, local o globalmente, a las propiedades físicas, en la medida que en el mundo natural, dos criaturas cualesquiera físicamente idénticas tendrán experiencias cualitativamente idénticas. Sin embargo, no es claro en absoluto que la conciencia sea lógicamente superveniente a propiedades físicas. Parece *lógicamente* posible, al menos para muchos, que una criatura físicamente idéntica a un ser consciente pueda no tener experiencias conscientes en absoluto, o que pueda tener experiencias conscientes de una clase diferente. (Algunos no están de acuerdo con esto, pero, por ahora, sólo utilizo la cuestión como ilustración.) Si esto es así, entonces la experiencia consciente superviene naturalmente pero no lógicamente a lo físico. La conexión necesaria entre la estructura física y la experiencia está asegurada sólo por las leyes de la naturaleza, y no por alguna fuerza lógica o conceptual.

La distinción entre superveniencia lógica y natural es vital para nuestros propósitos.[9] Podemos comprender de una manera intuitiva la distinción del siguiente modo. Si las propiedades B supervienen lógicamente a las propiedades A, entonces una vez que Dios (hipotéticamente) crea un mundo con ciertos hechos A, los hechos B los acompañan gratuitamente como consecuencia automática. Sin embargo, si las propiedades B poseen una mera superveniencia natural a las propiedades A, entonces, después de definir los hechos A, a Dios aún le resta trabajo por hacer para definir los hechos B: debe asegurarse de que exista una ley que relacione los hechos A con los hechos B. (Tomo prestada esta imagen de Kripke, 1972.) Una vez que esta ley existe, los hechos A relevantes producirán automáticamente los hechos B; pero podríamos, en principio, haber tenido una situación en la que esto no sucediese.

También, a veces, oímos hablar de la *superveniencia metafísica*, que no se basa ni en la necesidad lógica ni en la natural, sino en la "necesidad *tout court*", o "necesidad metafísica" como a veces se la conoce (inspirándose en la discusión sobre la necesidad *a posteriori* de Kripke (1972). Argumentaré más adelante que los mundos metafísicamente posibles son sólo los mundos lógicamente posibles (y que la posibilidad metafísica de los enunciados es posibilidad lógica con un giro semántico *a posteriori*), pero, por ahora, resultará inofensivo suponer que existe una noción de superveniencia metafísica, que debe ser especificada por analogía con las nociones de más arriba de superveniencia lógica y natural. También se menciona ocasionalmen-

te una noción de superveniencia "débil", pero parece ser demasiado débil como para expresar una relación de dependencia interesante entre propiedades.[10]

La distinción lógica-natural y la distinción global-local se entrecruzan. Es razonable hablar de superveniencia lógica global y de superveniencia lógica local, aunque nos ocuparemos más frecuentemente de la primera. Cuando hablo de superveniencia lógica sin otro modificador, me refiero a la superveniencia lógica global. También es coherente hablar de superveniencia natural global y local, pero las relaciones de superveniencia natural que nos interesan son, por lo general, locales o al menos localizables, por la simple razón de que la evidencia de una relación de superveniencia natural consiste usualmente en regularidades locales entre grupos de propiedades.[11]

## Un problema de la superveniencia lógica*

No podemos ignorar un problema técnico de la noción de superveniencia lógica. Este problema surge de la posibilidad lógica de un mundo físicamente idéntico al nuestro, pero que posea sustancia no física adicional que no existe en nuestro propio mundo: ángeles, ectoplasma y fantasmas, por ejemplo. Un mundo *concebible* exactamente como el nuestro, excepto que tiene algunos ángeles extra revoloteando en un dominio no físico hecho de ectoplasma. Si fuesen capaces de reproducirse y evolucionar estos ángeles podrían tener propiedades biológicas propias. Es posible que los ángeles pudieran tener todo tipo de creencias, y sus comunidades podrían tener una estructura social compleja.

El problema que estos ejemplos plantean es claro. El mundo de ángeles es físicamente idéntico al nuestro, pero es biológicamente distinto. Si el mundo de ángeles es lógicamente posible, entonces, de acuerdo con nuestra definición, las propiedades biológicas no son supervenientes a las propiedades físicas. Pero, *queremos* decir que las propiedades biológicas son supervenientes a las propiedades físicas, al menos en *este* mundo si no en el mundo de ángeles (¡suponiendo que no haya ángeles en el mundo real!). Intuitivamente, parece indeseable que la mera posibilidad lógica del mundo de ángeles obstaculice la determinación de las propiedades biológicas por las propiedades físicas en nuestro propio mundo.

Este tipo de problema hizo que algunos (por ejemplo, Haugeland, 1982; Petri, 1987) sugiriesen que la posibilidad y la necesidad lógicas son demasiado fuertes para servir como tipo relevante de posibilidad y necesidad en las relaciones de superveniencia, y que en su lugar debería utilizarse una variedad más débil como la posibilidad y la

necesidad natural. Pero esto haría que la distinción muy útil entre superveniencia lógica y natural que esbozamos más arriba resulte inútil, y también ignoraría el hecho de que hay un sentido muy real en el cual los hechos biológicos acerca de nuestro mundo están determinados lógicamente por los hechos físicos. Otros (por ejemplo, Teller, 1989) hicieron frente a la cuestión estipulando que los mundos con sustancias no física extra no son ni lógica ni metafísicamente posibles, a pesar de las apariencias, pero esto hace que la posibilidad lógica y metafísica parezca bastante arbitraria. Afortunadamente, estas medidas no son necesarias. Si formulamos la definición de un modo apropiado, es posible mantener una noción útil de superveniencia lógica compatible con la posibilidad de estos mundos.[12]

La clave de la solución es transformar la superveniencia en una tesis acerca de *nuestro* mundo (o más en general, acerca de mundos particulares). Esto concuerda con la intuición de que los hechos biológicos están lógicamente determinados por los hechos físicos en nuestro mundo, a pesar de la existencia de mundos bizarros en los que pueden no estarlo. De acuerdo con una definición revisada, las propiedades B son lógicamente supervenientes a las propiedades A si las propiedades B en nuestro mundo están lógicamente determinadas por las propiedades A en el siguiente sentido: en cualquier mundo posible con los mismos hechos A, ocurren los mismos hechos B.[13] La existencia de mundos con hechos B *extra* no contradice entonces la superveniencia lógica en nuestro mundo, siempre que *por lo menos* los hechos B que son verdaderos en nuestro mundo sean verdaderos en todos los mundos físicamente idénticos. Y esto será así por lo general (con una excepción que expondremos más abajo). Si hay un koala comiendo en un eucalipto en este mundo, habrá un koala idéntico comiendo en un eucalipto en cualquier mundo físicamente idéntico, independientemente de que ese mundo tenga ángeles merodeando.

Existe una complicación menor. Hay un cierto tipo de hechos biológicos acerca de nuestro mundo que no vale en el mundo de ángeles: el hecho de que el nuestro no tiene ectoplasma viviente, por ejemplo, y el hecho de que todos los seres vivientes se basan en el ADN. Quizá sea posible que el mundo de ángeles pueda incluso estar construido con ectoplasma causalmente dependiente de procesos físicos, de modo que una copulación de uombats en el plano físico podría producir a veces uombats ectoplasmáticos bebé en el plano no físico. Entonces, podría haber un uombat que no tenga hijos (en un cierto sentido) en nuestro mundo, con una contraparte que sí los tenga en el mundo angélico físicamente idéntico. Se deduce que la propiedad de no tener hijos no es superveniente según nuestra definición, y tampoco las propiedades al nivel del mundo como la de no poseer

ectoplasma viviente. No todos los hechos acerca del mundo surgen solamente de los hechos físicos.

Para analizar el problema, nótese que todos estos hechos involucran aseveraciones existenciales negativas y, por lo tanto, dependen no sólo de lo que sucede en nuestro mundo sino también de lo que no. No podemos esperar que estos hechos estén determinados por algún tipo de hechos localizados, ya que dependen no sólo de los sucesos locales en el mundo sino también de los límites del mismo. Las tesis de superveniencia sólo deberían aplicarse a los hechos y propiedades *positivos*, aquellos que no pueden ser negados simplemente agrandando un mundo. Podemos definir un hecho positivo en *W* como aquel que vale en todos los mundos que contienen a *W* como una parte propia;[14] una propiedad positiva es aquella que si está instanciada en un mundo *W*, también está instanciada por el individuo correspondiente en todos los mundos que contienen a *W* como parte propia.[15] La mayoría de los hechos y propiedades cotidianos son positivos, piénsese en la propiedad de ser un canguro, de tener un metro ochenta de altura o tener u hijo. Los hechos y propiedades negativos siempre involucrarán aseveraciones existenciales negativas de una forma u otra. Estas incluyen hechos existenciales explícitamente negativos tales como la no existencia del ectoplasma, hechos cuantificados universalmente tal como el hecho de que todos los seres vivientes se basan en el ADN, propiedades relacionales negativas como la de no tener hijos y superlativas como la propiedad de ser el organismo más fecundo en existencia.

En el futuro, deberá entenderse que las relaciones de superveniencia de las que nos ocuparemos están restringidas a hechos y propiedades positivos. Cuando afirmamos que las propiedades biológicas supervienen a las propiedades físicas, sólo son las propiedades biológicas positivas las que están en cuestión. Todas las propiedades de las que nos ocupamos son positivas —propiedades físicas locales y fenoménicas, por ejemplo— de modo que esta no es una gran restricción.

Por lo tanto, la definición de la superveniencia lógica global de las propiedades B a las propiedades A se reduce a lo siguiente: para cualquier mundo lógicamente posible *W* que es A-indiscernible de nuestro mundo, los hechos B verdaderos en nuestro mundo son también verdaderos en *W*. No necesitamos incorporar ninguna cláusula acerca de la positividad, pero usualmente supondremos que los únicos hechos B y propiedades B relevantes son hechos y propiedades positivos. De modo similar, las propiedades B supervienen local y lógicamente a las propiedades A cuando para cada individuo concreto *x* y todo individuo lógicamente posible *y*, si *y* es A-indiscernible de *x*,

entonces las propiedades B instanciadas por $x$ son instanciadas por $y$. Más breve y generalmente: las propiedades B supervienen lógicamente a las propiedades A si los hechos B acerca de situaciones reales están implicados por los hechos A, donde las situaciones se interpretan como mundos e individuos en los casos global y local respectivamente. Esta definición captura la idea de que las aseveraciones de superveniencia son por lo general aseveraciones acerca de nuestro mundo, pero conserva también el papel clave de la necesidad lógica.[16]

## Superveniencia y materialismo

Las superveniencias lógica y natural tienen ramificaciones bastante diferentes para la ontología: esto es, para la cuestión de qué es lo que existe en el mundo. Si las propiedades B son lógicamente supervenientes a las propiedades A, entonces existe un sentido en el cual una vez que los hechos A están definidos, los hechos B ocurren sin costo adicional. Una vez que Dios (hipotéticamente) hizo que todos los hechos físicos en nuestro mundo fueran como son, los hechos biológicos los acompañaron sin ningún costo ulterior. Los hechos B meramente redescriben lo que los hechos A describen. Pueden ser hechos *diferentes* (un hecho acerca de elefantes no es un hecho microfísico), pero no son hechos *ulteriores*.

Con la mera superveniencia natural, la ontología no es tan directa. Las conexiones legaliformes contingentes conectan diferentes características del mundo. En general, si las propiedades B son supervenientes en forma simplemente natural a las propiedades A en nuestro mundo, entonces *podría* existir un mundo en el que nuestros hechos A fuesen válidos sin los hechos B. Como vimos antes, una vez que Dios fijó todos los hechos A, para fijar los hechos B tuvo que hacer más trabajo. Los hechos B van más allá de los hechos A y su satisfacción implica que hay algo nuevo en el mundo.

Con esto en mente podemos formular con precisión la doctrina ampliamente aceptada del *materialismo* (o *fisicalismo*), que, por lo general, se interpreta que sostiene que todo en el mundo es físico, o que no hay nada más allá de lo físico, o que los hechos físicos en cierto sentido agotan todos los hechos acerca del mundo. En nuestro lenguaje, el materialismo es verdadero si todos los hechos positivos acerca del mundo son lógicamente supervenientes en forma global a los hechos físicos. Esto captura la noción intuitiva de que si el materialismo es verdadero, entonces una vez que Dios determinó los hechos físicos acerca del mundo, todos los hechos quedaron determinados.

(O, al menos, todos los hechos positivos quedaron determinados. La restricción a hechos positivos es necesaria para asegurar que los

mundos con hechos ectoplasmáticos extra no atenten en contra del materialismo en nuestro mundo. Los hechos existenciales negativos como "No hay ángeles" no son en sentido estricto lógicamente supervenientes a lo físico, pero su no superveniencia es compatible con el materialismo. En cierto sentido, para determinar los hechos negativos, Dios tuvo que hacer algo más que determinar los hechos físicos; tuvo que declarar "Eso es todo". (Si quisiéramos, podríamos agregar un hecho de segundo orden "Eso es todo" a la base de superveniencia en la definición del materialismo, en cuyo caso la restricción a los hechos positivos podría eliminarse.)

Según esta definición, el materialismo es verdadero si todos los hechos positivos acerca de nuestro mundo están implicados por los hechos físicos.[17] Esto es, el materialismo es verdadero si para cualquier mundo lógicamente posible W que es físicamente indiscernible de nuestro mundo, todos los hechos positivos verdaderos en nuestro mundo son verdaderos en W. Esto es equivalente, a su vez, a la tesis de que cualquier mundo que es físicamente indiscernible del nuestro contiene una copia de nuestro mundo como una parte (propia o impropia), lo que parece ser una definición intuitivamente correcta.[18] (Esta es equivalente a la definición de fisicalismo dada por Jackson [1994], cuyo criterio es que todo duplicado físico mínimo de nuestro mundo es un duplicado *simpliciter* de él.)[19]

Discutiré esta cuestión más en profundidad en el capítulo 4, donde justificaré más extensamente esta definición de materialismo. Algunos podrían objetar el uso de la posibilidad lógica en lugar de la posibilidad *tout court* o "posibilidad metafísica". Estas personas pueden sustituir posibilidad metafísica por posibilidad lógica en la definición de más arriba. Luego, argumentaré que significan lo mismo.

## 2. La explicación reductiva

El notable progreso de la ciencia durante los últimos siglos nos ha dado buenas razones para creer que hay muy poco que sea totalmente misterioso acerca del mundo. Para casi todo fenómeno natural por encima del nivel de la física microscópica, parece existir en principio una *explicación reductiva*: esto es, una explicación completa en términos de entidades más simples. En estos casos, cuando formulamos una concepción apropiada de los procesos de nivel inferior, la explicación de los fenómenos de nivel superior surge automáticamente.

Los fenómenos biológicos suministran una clara ilustración. La reproducción puede explicarse formulando una concepción de los

mecanismos genéticos y celulares que les permiten a los organismos producir otros organismos. La adaptación puede explicarse formulando una concepción de los mecanismos que llevan a cambios apropiados en el funcionamiento externo en respuesta a la estimulación ambiental. La vida misma se explica a través de los diversos mecanismos que producen la reproducción, la adaptación y otros fenómenos similares. Una vez que contamos la historia de nivel inferior con suficiente detalle, cualquier sentido de misterio fundamental desaparece: los fenómenos que debían ser explicados lo han sido.

Se puede contar una historia similar sobre la mayoría de los fenómenos naturales. En física, explicamos el calor contando una historia apropiada acerca de la energía y la excitación de las moléculas. En astronomía, explicamos las fases de la luna entrando en los detalles del movimiento orbital y la reflexión óptica. En geofísica, los terremotos se explican por medio de una descripción de la interacción de las masas subterráneas. En la ciencia cognitiva, para explicar un fenómeno como el aprendizaje, todo lo que tenemos que hacer es, en una primera aproximación (dejando de lado cualquier preocupación sobre la *experiencia* del aprendizaje), explicar diversos mecanismos funcionales, los mecanismos que dan origen a los cambios apropiados en la conducta en respuesta a la estimulación ambiental. Muchos de los detalles de estas explicaciones se nos escapan en la actualidad, y es probable que resulten muy complejos, pero sabemos que si indagamos lo suficiente sobre la historia de bajo nivel, luego la historia de alto nivel la acompañará.

No definiré en forma precisa la noción de explicación reductiva sino hasta más adelante. Por ahora, deberá permanecer caracterizada por los ejemplos que dimos. Sin embargo, puedo formular algunas advertencias acerca de lo que la explicación reductiva no es. Una explicación reductiva de un fenómeno no exige una *reducción* de ese fenómeno, al menos en algunos sentidos de este ambiguo término. En cierto modo, los fenómenos que pueden realizarse en muchos sustratos físicos diferentes —el aprendizaje, por ejemplo— podrían no ser reducibles en el sentido de que no podemos *identificar* el aprendizaje con ningún fenómeno específico de nivel inferior. Pero esta múltiple realizabilidad no impide *explicar* reductivamente cualquier instancia de aprendizaje en términos de fenómenos de nivel inferior.[20] La explicación reductiva de un fenómeno tampoco debería confundirse con una reducción de una *teoría* de alto nivel. A veces una explicación reductiva de un fenómeno proporcionará una reducción de una teoría de alto nivel preexistente, pero otras veces mostrará que esas teorías están en el camino erróneo. Con frecuencia podría no existir una teoría de alto nivel para reducir.

La explicación reductiva no es la finalidad última de toda explicación. Hay muchos otros tipos de explicaciones, algunas de las cuales pueden arrojar más luz sobre un fenómeno que una reductiva en una instancia dada. Existen explicaciones *históricas*, por ejemplo, que dan cuenta de la génesis de un fenómeno como la vida, mientras que una explicación reductiva sólo ofrece una concepción sincrónica de cómo funcionan los sistemas vivientes. Existe también todo tipo de explicaciones de *alto nivel*, tal como la de aspectos de la conducta en términos de creencias y deseos. Aun cuando esta conducta podría en principio ser explicable en forma reductiva, una explicación de nivel superior es con frecuencia más comprensible y esclarecedora. No debería considerarse que las explicaciones reductivas desplazan a estos otros tipos de explicación. Cada una tiene su lugar.

## La explicación reductiva por medio del análisis funcional

¿Qué es lo que permite que fenómenos tan diversos como la reproducción, el aprendizaje y el calor sean explicados reductivamente? En todos estos casos, la naturaleza de los conceptos requeridos para caracterizar los fenómenos es crucial. Si alguien pusiese objeciones a una explicación celular de la reproducción diciendo, "Esto explica cómo un proceso celular puede llevar a la producción de una entidad física compleja que es similar a la entidad original, pero no explica la *reproducción*", tendríamos poca paciencia, porque eso es todo lo que "reproducción" *significa*. En general, una explicación reductiva de un fenómeno está acompañada por algún *análisis* aproximado, implícito o explícito, del fenómeno en cuestión. La noción de reproducción puede analizarse aproximadamente en términos de la capacidad de un organismo para producir de una cierta manera otro organismo. Se deduce que una vez que explicamos el proceso por el cual un organismo produce otro organismo, hemos explicado esa instancia de reproducción.

El punto podría parecer trivial, pero la posibilidad de este tipo de análisis apuntala la posibilidad de la explicación reductiva en general. Sin un análisis de este tipo, no habría ningún puente explicativo desde los hechos físicos de bajo nivel al fenómeno en cuestión. Con él, todo lo que necesitamos es mostrar cómo ciertos mecanismos físicos de nivel inferior permiten que el análisis se satisfaga, y esto resulte en una explicación.

Para los fenómenos más interesantes que requieren explicación, incluyendo fenómenos como la reproducción y el aprendizaje, las nociones relevantes suelen analizarse *funcionalmente*. El núcleo de estas nociones puede caracterizarse en términos del desempeño de alguna función o funciones (donde "función" se interpreta de modo causal, no teleológico), o en términos de la capacidad de realizar

esas funciones. Se deduce que una vez que explicamos cómo estas se realizan, entonces hemos explicado el fenómeno en cuestión. Una vez que explicamos cómo un organismo desempeña la función de producir otro organismo, hemos explicado la reproducción, porque todo lo que significa reproducir es realizar esa función. Lo mismo vale para una explicación del aprendizaje. Lo que significa que un organismo aprenda es, aproximadamente, que sus capacidades conductuales se adapten en forma apropiada en respuesta a la estimulación ambiental. Si explicamos cómo el organismo es capaz de realizar las funciones relevantes, entonces hemos explicado el aprendizaje.

(Cuanto más, podemos haber fallado en explicar los aspectos *fenoménicos* del aprendizaje, que dejo de lado aquí por razones obvias. Si existe un elemento fenoménico en el concepto de aprendizaje, entonces esa parte del fenómeno puede quedar inexplicada; pero aquí me concentro en los aspectos psicológicos del aprendizaje, que constituyen el núcleo del concepto.)

Explicar el desempeño de esas funciones es, en principio, bastante simple. Si los resultados de dichas funciones son caracterizables físicamente, y si todos los sucesos físicos tienen causas físicas, entonces debería haber una explicación física para el desempeño de una función de este tipo. Sólo debemos mostrar cómo ciertos tipos de estados son responsables de la producción de los estados resultantes apropiados, mediante un proceso causal de acuerdo con las leyes de la naturaleza. Por supuesto, los detalles de este tipo de explicación física pueden ser no triviales. Ciertamente, los detalles constituyen una gran parte de cualquier explicación reductiva, mientras que el componente de análisis con frecuencia es trivial. Pero, una vez que se tienen los detalles relevantes, una historia sobre la causalidad física de bajo nivel explicará cómo se realizan las funciones relevantes y, por lo tanto, explicará el fenómeno en cuestión.

Inclusive una noción física como el calor puede interpretarse de una manera funcional: aproximadamente, el calor es el tipo de cosa que expande los metales, es causado por el fuego, lleva a un tipo particular de sensación, etc. Una vez que tenemos una concepción de cómo estas diversas relaciones causales se cumplen, entonces tenemos una concepción del calor. El calor es un *concepto basado en el papel causal*, es decir, que se caracteriza en términos de aquellas cosas que típicamente lo causan y de aquellas a las que típicamente causa, bajo condiciones apropiadas. Una vez que la investigación empírica muestra cómo se realiza el papel causal relevante, el fenómeno ha sido explicado.

Aquí existen algunas complicaciones técnicas, pero no son esenciales. Por ejemplo, Kripke (1980) señaló una diferencia entre un

término como "calor" y la descripción asociada de un papel causal: dado que el calor ocurre debido al movimiento de las moléculas, entonces este movimiento podría calificar como calor en un mundo contrafáctico, independientemente de que estas moléculas tengan un papel causal relevante. De todas maneras, sigue siendo válido que *explicar* el calor involucra explicar el desempeño del papel causal, no el movimiento de las moléculas. Para ver esto, nótese que la equivalencia del calor con el movimiento de las moléculas se conoce *a posteriori*: sabemos esto *como resultado* de explicar el calor. El concepto de calor que teníamos *a priori* —antes de que el fenómeno fuese explicado— era aproximadamente el de "la cosa que tiene este papel causal en el mundo real". Una vez que descubrimos cómo se realiza el papel causal, tenemos una explicación del fenómeno. Como bonificación, sabemos qué *es* el calor. Es el movimiento de las moléculas, ya que este movimiento es lo que realiza el papel causal relevante en el mundo real.

Una segunda complicación menor es que muchos conceptos basados en el papel causal son algo ambiguos entre el estado que desempeña un cierto papel causal y el desempeño concreto de ese papel. Puede interpretarse que "calor" denota las moléculas que hacen el trabajo causal o el propio proceso causal (calentamiento). De modo similar, "percepción" puede usarse para referir al acto de percibir o al estado interno que surge como resultado. Sin embargo, nada importante gira en torno de esta ambigüedad. Una explicación de cómo se realiza el papel causal explicará el calor o la percepción en cualquiera de estos sentidos.

Una tercera complicación es que muchos conceptos basados en el papel causal se caracterizan parcialmente en términos de su efecto sobre la *experiencia*: por ejemplo, el calor se interpreta en forma natural como la causa de las sensaciones de calor. ¿Significa esto que tenemos que explicar las sensaciones de calor antes de que podamos explicar el calor? Por supuesto, no poseemos ninguna buena concepción de las sensaciones de calor (o de la experiencia en general), de modo que lo que ocurre en la práctica es que esa parte del fenómeno queda inexplicada. Si podemos explicar cómo se produce el movimiento molecular en ciertas condiciones y causa que los metales se expandan y estimula nuestra piel de ciertos modos, entonces la observación de que este movimiento está *correlacionado* con sensaciones de calor es suficientemente buena. De la correlación, inferimos que existe casi seguramente una conexión causal. Sin duda, ninguna explicación del calor estará completa hasta que no tengamos una concepción de cómo funciona esa conexión causal, pero la concepción incompleta es suficientemente buena para la mayoría de los

propósitos. En cierta forma es paradójico que terminemos explicando casi todo acerca de un *fenómeno* excepto los detalles de cómo este afecta nuestra fenomenología, pero no es un problema en la práctica. No sería un estado de cosas feliz si tuviésemos que poner en suspenso al resto de la ciencia hasta que tuviéramos una teoría de la conciencia.

## Las explicaciones reductivas en la ciencia cognitiva

El paradigma de la explicación reductiva mediante el análisis funcional se desempeña perfectamente en la mayoría de las áreas de la ciencia cognitiva, al menos en principio. Como vimos en el capítulo anterior, la mayor parte de los conceptos mentales no fenoménicos pueden analizarse funcionalmente. Los estados psicológicos son caracterizables en términos del papel causal que desempeñan. Para explicar esos estados, explicamos cómo se realiza la causación pertinente.

En principio, se puede hacer esto formulando una concepción de la neurofisiología subyacente. Si explicamos cómo ciertos estados neurofisológicos son responsables de la realización de las funciones en cuestión, entonces hemos explicado el estado psicológico. Sin embargo, no siempre necesitamos descender al nivel neurofisiológico. Frecuentemente podemos explicar algún aspecto de la mentalidad exhibiendo un *modelo cognitivo* apropiado, esto es, exhibiendo los detalles de la organización causal abstracta de un sistema cuyos mecanismos son suficientes para realizar las funciones pertinentes, sin especificar el sustrato fisicoquímico en el cual esta organización causal está implementada. De este modo, damos una explicación *posiblemente cómo* sobre un aspecto determinado de la psicología, en el sentido de que mostramos cómo los mecanismos causales apropiados podrían realizar los procesos mentales pertinentes. Si estamos interesados en explicar los estados mentales de un organismo o tipo de organismo *real* (por ejemplo, el aprendizaje en los seres humanos, en lugar de la posibilidad del aprendizaje en general), este tipo de explicación debe complementarse con una demostración de que la organización causal del modelo refleja la organización causal del organismo en cuestión.

Para explicar la posibilidad del aprendizaje, podemos exhibir un modelo cuyos mecanismos llevan a los cambios apropiados en la capacidad conductual en respuesta a diversos tipos de estimulación ambiental; por ejemplo, un modelo de aprendizaje conexionista. Para explicar el aprendizaje humano, debemos también mostrar que un modelo de este tipo refleja la organización causal responsable del

desempeño de dichas funciones en los seres humanos. El segundo paso es usualmente difícil: no podemos exhibir una correspondencia de este tipo en forma directa, debido a nuestra ignorancia de la neurofisiología, de modo que por lo general tenemos que buscar evidencia indirecta, como similitudes cualitativas en patrones de respuesta, mediciones de tiempos y otras cosas similares. Esta es una razón de por qué la ciencia cognitiva está actualmente en un estado de subdesarrollo. Pero, como es usual, la posibilidad en principio de una explicación de esta clase es una consecuencia directa de la naturaleza funcional de los conceptos psicológicos.

Desafortunadamente, el tipo de explicación funcional que tiene tan buen desempeño en los estados psicológicos no parece funcionar para explicar estados fenoménicos. La razón es simple. Cualquiera sea la concepción funcional de la cognición humana que formulemos, siempre queda una *pregunta ulterior*: ¿Por qué este tipo de funcionamiento está acompañado por la conciencia? Esta pregunta no surge para los estados psicológicos. Si preguntamos acerca de un modelo funcional particular de aprendizaje, "¿Por qué este funcionamiento está acompañado de aprendizaje?", la respuesta apropiada es una respuesta semántica: "Porque lo que *significa* aprender es funcionar de esa forma". No existe un análisis correspondiente del concepto de conciencia. Los estados fenoménicos, a diferencia de los estados psicológicos, no se definen por los papeles causales que desempeñan. Se deduce que la explicación de cómo se realiza algún papel causal no es suficiente para explicar la conciencia. Después de que hemos explicado la realización de una función determinada, el hecho de que la conciencia acompañe el desempeño de la función (si lo hace) permanece sin explicación.

Podemos decirlo del siguiente modo. Dada una concepción funcional apropiada del aprendizaje, es *lógicamente imposible* que algo pueda instanciar esa concepción sin aprender (excepto, quizás, si el aprendizaje requiriese de la conciencia). Sin embargo, independientemente de la concepción funcional de la cognición que formulemos, parece lógicamente posible que esa concepción pueda instanciarse sin ninguna conciencia acompañante. Podría ser naturalmente imposible —la conciencia podría *surgir* de hecho de esa organización funcional en el mundo real— pero lo importante es que la noción es lógicamente coherente.

Si esto es lógicamente posible, entonces cualquier concepción funcional y, por cierto, cualquier concepción física de los fenómenos mentales será básicamente incompleta. Para utilizar una frase acuñada por Levine (1983), existe una *brecha explicativa* entre dichas concepciones y la propia conciencia. Aun si la organización funcional

apropiada siempre diese origen, en la práctica, a la conciencia, la cuestión de *por qué* esto es así permanece sin respuesta. Este punto será desarrollado en forma detallada más adelante.

Si esto es así, se deduce que existirá una brecha explicativa parcial en cualquier concepto mental que tenga un elemento fenoménico. Si se requiere la experiencia consciente para la creencia o el aprendizaje, por ejemplo, podríamos no tener una explicación completamente reductiva de estos fenómenos. Pero, al menos, tenemos razones para creer que los aspectos *psicológicos* de estas características mentales —que están supuestamente en el centro de los conceptos relevantes— podrán, en principio, explicarse reductivamente. Si dejamos de lado las preocupaciones por la fenomenología, la ciencia cognitiva parece tener los recursos necesarios para hacer un buen trabajo en la explicación de la mente.

## 3. Superveniencia lógica y explicación reductiva

La epistemología de la explicación reductiva satisface la metafísica de la superveniencia de un modo directo. Un fenómeno natural es explicable de un modo reductivo en términos de algunas propiedades de bajo nivel precisamente cuando es lógicamente superveniente a esas propiedades. Es explicable de un modo reductivo en términos de propiedades físicas —o simplemente "explicable reductivamente"— cuando es lógicamente superveniente a lo físico.

Para decirlo con más cuidado: un fenómeno natural es reductivamente explicable en términos de algunas propiedades de más bajo nivel si la propiedad de instanciar ese fenómeno es en forma global lógicamente superveniente a las propiedades de bajo nivel mencionadas. Un fenómeno es explicable reductivamente *simpliciter* si la propiedad de ejemplificar ese fenómeno es en forma global lógicamente superveniente a propiedades físicas.

Esto puede interpretarse como una *explicitación* de la noción de explicación reductiva, quizá con algún elemento de estipulación. Debería resultar claro del análisis anterior que nuestra noción previa de explicación reductiva implica la superveniencia lógica (global). Si la propiedad de ejemplificar un fenómeno falla en supervenir lógicamente a algunas propiedades de nivel inferior, entonces para cualquier concepción de nivel inferior de esas propiedades siempre habrá una pregunta ulterior sin respuesta: ¿por qué este proceso de nivel inferior está acompañado por el fenómeno? La explicación reductiva requiere de algún tipo de análisis del fenómeno en cuestión, donde los hechos de bajo nivel implican la realización del análisis. De este modo la explicación reductiva requiere una relación de superveniencia

lógica. Por ejemplo, es precisamente porque la reproducción es lógicamente superveniente a hechos de nivel inferior que es explicable de modo reductivo en términos de esos hechos.

Resulta algo menos claro que la superveniencia lógica *sea suficiente* para la explicabilidad reductiva. Si un fenómeno *P* superviene lógicamente a algunas propiedades de nivel inferior, entonces dada una concepción de los hechos de nivel inferior asociados con una instancia de *P*, la ejemplificación de *P* es una consecuencia lógica. Una concepción de los hechos de nivel inferior producirá por lo tanto automáticamente una explicación de *P*. No obstante, una explicación de esta clase puede, a veces, parecer insatisfactoria por dos razones. Primero, los hechos de nivel inferior podrían ser un vasto fárrago de detalles aparentemente arbitrarios sin ninguna unidad explicativa clara. Una descripción de todos los movimientos moleculares subyacentes a una instancia de aprendizaje podría ser un ejemplo de esta clase. Segundo, es posible que diferentes instancias de *P* puedan estar acompañadas de conjuntos muy diferentes de hechos de bajo nivel, de modo que las explicaciones de instancias particulares no produzcan una explicación del fenómeno como tipo.

Una opción es sostener que la superveniencia lógica es *necesaria* para la explicación reductiva, pero no suficiente. Esto es todo lo que se requiere para mis argumentos acerca de la conciencia en el próximo capítulo. Pero es más provechoso notar que existe *una* noción útil de explicación reductiva tal que la superveniencia lógica es necesaria y suficiente. En lugar de interpretar que los problemas de más arriba indican que las concepciones en cuestión no son *explicaciones*, podemos en cambio interpretar que indican que una explicación reductiva no es necesariamente una explicación *esclarecedora*. Más bien, una explicación reductiva es una explicación que *elimina el misterio*.

Como hice notar con anterioridad, la explicación reductiva no es el fin último de la explicación. Su papel principal es remover cualquier sentido profundo de misterio que rodee a un fenómeno de alto nivel. Lo hace reduciendo la primitividad y arbitrariedad del fenómeno en cuestión a la primitividad y arbitrariedad de procesos de nivel inferior. En la medida en que los procesos de bajo nivel pueden ellos mismos ser relativamente primitivos y arbitrarios, es posible que una explicación reductiva no nos dé una comprensión *profunda* de un fenómeno, pero al menos elimina cualquier sensación de que algo "extra" ocurre.

Sin embargo, la brecha entre una explicación reductiva y una explicación esclarecedora puede, por lo general, cerrarse mucho más que esto. Ello se debe a dos hechos básicos sobre la física de nuestro

mundo: la *autonomía* y la *simpleza*. La causalidad microfísica y la explicación parecen ser autónomas, en el sentido de que todo suceso físico tiene una explicación física; las leyes de la física son suficientes para explicar los sucesos de la física en sus propios términos. Más aún, las leyes en cuestión son razonablemente simples, de modo que las explicaciones de las que hablamos son de algún modo compactas. Ambas cosas podrían haber sido de otra manera. Podríamos haber vivido en un mundo en el que leyes fundamentales, primitivas, emergentes gobernasen la conducta de configuraciones de alto nivel como los organismos, con una causalidad descendente asociada que domina sobre cualquier ley microfísica relevante. (Los emergentistas ingleses como Alexander [1920] y Broad [1925] creían que nuestro mundo era de esta clase.) Alternativamente, nuestro mundo podría haber sido un mundo en el cual la conducta de las entidades microfísicas estuviera gobernada por un vasto conjunto de leyes intrincadas, o quizás un mundo en el cual la conducta microfísica fuera anárquica y caótica. En mundos como estos, habría poca esperanza de lograr una explicación reductiva esclarecedora, ya que la primitividad de las concepciones de bajo nivel no podrían nunca llegar a simplificarse.

Pero el mundo real, con su autonomía y simplicidad de bajo nivel, parece por lo general permitir que se extraiga de él un sentido, incluso en el caso de los procesos complejos. Los hechos de bajo nivel subyacentes a los fenómenos de alto nivel tienen una unidad básica que hace posible una explicación comprensible. Dada una instancia causal de alto nivel, tal como la liberación de un gatillo que hace que un arma de fuego se dispare, podemos no sólo aislar un manojo de hechos de nivel inferior que determinan esta causa; también podemos contar una historia relativamente simple de cómo la causa se hace posible, encapsulando esos hechos bajo ciertos principios simples. Esto no siempre funciona. Puede ocurrir que algunos dominios, como los de la sociología y la economía, estén tan alejados de la simpleza de los procesos de bajo nivel que resulte imposible una explicación reductiva esclarecedora, aunque los fenómenos sean lógicamente supervenientes. Si esto es así, que así sea: podemos contentarnos con explicaciones de alto nivel en estos dominios, notando al mismo tiempo que la superveniencia lógica implica que en principio existe una explicación reductiva, aunque quizás una que sólo un superser podría comprender.

Adviértase también que según esta concepción la explicación reductiva es fundamentalmente *particular*, da cuenta de instancias particulares de un fenómeno, sin necesariamente dar cuenta de todas las instancias juntas. Esto es lo que deberíamos esperar. Si una propiedad puede ser instanciada de muchos modos diferentes, no

podemos esperar que una sola explicación cubra todas las instancias. La temperatura se instancia en formas bastante diferentes en distintos medios, por ejemplo, y existen distintas explicaciones para cada una. En un nivel mucho más alto, es muy improbable que exista una sola explicación que cubra todas las instancias de asesinato. Sin embargo, por lo general existe una cierta unidad a través de las explicaciones de particulares, en el sentido de que una buena explicación de uno es con frecuencia una explicación de muchos. Esto es, nuevamente, una consecuencia de la simplicidad subyacente de nuestro mundo, más que una propiedad necesaria de la explicación. En nuestro mundo, las historias unificadoras simples que podemos contar acerca de los procesos de nivel inferior suelen aplicarse de manera generalizada o, al menos, a través de un amplio espectro de casos particulares. También suele suceder, en especial en las ciencias biológicas, que los casos particulares tienen una ascendencia común que lleva a una similitud en los procesos de bajo nivel involucrados. De este modo, el segundo problema mencionado, el de unificar las explicaciones de las instancias específicas de un fenómeno, no es un problema tan significativo como podría parecer. En cualquier caso, lo fundamental es la explicación de casos particulares.

Hay mucho más que podría decirse acerca de cerrar la brecha entre la explicación reductiva y la explicación esclarecedora, pero la cuestión merece un tratamiento detallado por derecho propio y no resulta fundamental para mis propósitos. Lo que es más importante es que si no existe una superveniencia lógica (como argumentaré que ocurre para la conciencia), entonces *cualquier* clase de explicación reductiva fracasa, aun si somos generosos acerca de lo que puede considerarse una explicación. También es importante notar que la superveniencia lógica elimina cualquier misterio *metafísico* residual acerca de un fenómeno de alto nivel, debido a que reduce cualquier primitividad en ese fenómeno a la primitividad de los hechos de nivel inferior. Resulta de importancia secundaria que si la superveniencia lógica es válida, entonces es posible algún tipo de explicación reductiva. Aunque esta clase de explicaciones puede fallar en ser esclarecedora o útil, este fracaso no es ni de cerca tan fundamental como el fracaso de la explicación en dominios en los que la superveniencia lógica no es válida.

### Notas adicionales sobre la explicación reductiva

Unas pocas notas adicionales. Primero, una explicación reductiva práctica de un fenómeno por lo general no llega hasta el nivel microfísico. Hacerlo así sería enormemente difícil ya que daría lugar

a todos los problemas de primitividad recién mencionados. Los fenómenos de alto nivel, en cambio, se explican en términos de algunas propiedades de un nivel ligeramente más básico, como cuando se explica la reproducción en términos de mecanismos celulares, o las fases de la luna como movimientos orbitales. A su vez, esperamos que los fenómenos más básicos sean reductivamente explicables como algo todavía más importante. Si todo sale bien, los fenómenos biológicos pueden ser explicables en términos de fenómenos celulares, que son explicables en términos de fenómenos bioquímicos, que son explicables en términos de fenómenos químicos, que son explicables en términos de fenómenos físicos. En lo que respecta a los fenómenos físicos, intentamos unificarlos lo más posible, pero en algún nivel la física debe tomarse como primitiva: podría no haber una explicación de por qué las leyes fundamentales o las condiciones fronterizas son como son. Por el momento, esta escalera explicativa es poco más que una quimera, pero se han realizado progresos significativos. Dada la superveniencia lógica, junto con la simplicidad y la autonomía del nivel más bajo, este tipo de conexión explicativa entre las ciencias debería ser posible en principio. Es un problema abierto si las complejidades de la realidad pueden hacer que esto no sea factible en la práctica.

Segundo, es al menos concebible que un fenómeno pueda ser reductivamente explicable en términos de propiedades de nivel inferior sin que sea reductivamente explicable *simpliciter*. Esto podría ocurrir en una situación en la que ciertas propiedades C sean lógicamente supervenientes a las propiedades B, y por lo tanto explicables en términos de las propiedades B, pero donde estas últimas no sean lógicamente supervenientes a lo físico. Hay claramente un sentido en el que una explicación de esta clase es reductiva y otro sentido en el que no lo es. En su mayor parte, me ocuparé de la explicación reductiva en términos de lo físico o en términos de propiedades que son explicables en términos de lo físico. Aunque las propiedades C aquí sean reductivamente explicables en un sentido relativo, su propia existencia entraña el fracaso de la explicación reductiva en general.

Tercero, la superveniencia lógica *local* es un requerimiento demasiado estricto para la explicación reductiva. Podemos también explicar reductivamente las propiedades dependientes del contexto de un individuo mediante una formulación de una concepción de cómo las relaciones ambientales relevantes llegan a satisfacerse. Si un fenómeno es globalmente superveniente, será reductivamente explicable en términos de algunos hechos de nivel inferior, aun cuando estos se encuentren muy dispersos en el espacio y el tiempo.

Cuarto, en principio, existen dos proyectos en la explicación reductiva de fenómenos como la vida, el aprendizaje, el calor. Primero tenemos un proyecto de *explicitación*, en el que clarificamos exactamente qué es lo que debe explicarse por medio del análisis. Por ejemplo, el aprendizaje podría analizarse como una cierta clase de proceso adaptativo. Segundo, existe un proyecto de *explicación*, en el que vemos cómo ese análisis se realiza en términos de hechos de bajo nivel. El primer proyecto es conceptual, y el segundo empírico. Para muchos o la mayoría de los fenómenos, la etapa conceptual será bastante trivial. Sin embargo, para algunos fenómenos como la creencia la explicitación puede ser un obstáculo importante en sí mismo. En la práctica, por supuesto, no existe una separación nítida entre los proyectos, ya que la explicitación y la explicación ocurren en paralelo.

## 4. Verdad conceptual y verdad necesaria*

En mi concepción de la superveniencia y la explicación, me he apoyado fuertemente en las nociones de posibilidad y necesidad lógicas. Es tiempo de decir algo más acerca de esto. El modo básico de comprender la necesidad lógica de un enunciado es en términos de su verdad a través de todos los mundos lógicamente posibles. Se debe tener un cierto cuidado en darle sentido a la clase relevante de mundos y al modo como los enunciados se evalúan en ellos; más adelante en este apartado analizaré esta cuestión más detenidamente. También es posible explicitar la necesidad lógica de un enunciado a partir de su valor de verdad en virtud de su significado: un enunciado es lógicamente necesario si su verdad está asegurada por el significado de los conceptos involucrados. Pero, nuevamente, se requiere un cierto cuidado para comprender exactamente cómo deberían interpretarse los "significados". Analizaré estos dos modos de ver las cosas y su relación, más adelante, en este apartado.

(Como antes, la noción de necesidad lógica no debe identificarse con una noción más estrecha que involucra la derivabilidad en la lógica de primer orden o en algún otro formalismo sintáctico. Podría argumentarse que la justificación de los axiomas y las reglas en estos formalismos depende precisamente de su necesidad lógica en el sentido más amplio y primitivo.)

Todo esto requiere que se tome seriamente, al menos en cierta medida, la noción de *verdad conceptual*, esto es, la noción de que algunos enunciados son verdaderos o falsos simplemente en virtud de los significados de los términos involucrados. Algunos elementos clave en mi exposición hasta el momento dependieron de carac-

terizaciones de diversos conceptos. Por ejemplo, ofrecí una concepción de la explicación reductiva de la reproducción argumentando que los detalles de bajo nivel implican que se realizan ciertas funciones y que la realización de estas es todo lo que hay en el concepto de reproducción.

La noción de verdad conceptual ha tenido mala reputación en algunos círculos desde la crítica de Quine (1951); este filósofo sostuvo que no existe ninguna distinción útil entre las verdades conceptuales y las verdades empíricas. Las objeciones a estas nociones se agrupan por lo general en torno de los siguientes puntos:

1. La mayor parte de los conceptos no tienen definiciones que enuncien condiciones necesarias y suficientes (esta observación fue hecha un número de veces pero con frecuencia se asocia a Wittgenstein, 1953).

2. La mayoría de las aparentes verdades conceptuales son de hecho revisables, y podrían ser removidas ante suficiente evidencia empírica (un punto planteado por Quine).

3. Las consideraciones acerca de la necesidad *a posteriori*, bosquejadas por Kripke (1972), muestran que las condiciones de aplicación de muchos términos a través de los mundos posibles no pueden conocerse *a priori*.

Estas consideraciones van en contra de un enfoque excesivamente simplista de la verdad conceptual, pero no en contra del modo como utilizo estas nociones. En particular, resulta que la clase de *condicionales de superveniencia* —"Si los hechos A acerca de una situación son *X*, entonces los hechos B son *Y*", donde los hechos A especifican completamente una situación en un nivel fundamental— no son afectados por estas consideraciones. Estas son las únicas verdades conceptuales que mis argumentos necesitan, y veremos que ninguna de las consideraciones de más arriba los contradicen. También analizaremos más en detalle la relación entre la verdad conceptual y la verdad necesaria, y especificaremos su papel en la comprensión de la superveniencia lógica.

## Definiciones

La ausencia de definiciones precisas es la menos seria de las dificultades de la verdad conceptual. Ninguno de mis argumentos depende de la existencia de definiciones de este tipo. Ocasionalmente me apoyaré en el análisis de varias nociones, pero estos análisis sólo necesitan ser aproximados, sin ninguna pretensión de proporcionar

condiciones necesarias y suficientes precisas. La mayoría de los conceptos (por ejemplo, "vida") son algo vagos en su aplicación, y tiene poco sentido intentar remover esa vaguedad mediante una precisión arbitraria. En lugar de decir "Un sistema está vivo si y sólo si se reproduce, se adapta con una utilidad de 800 o más, y metaboliza con una eficiencia del 75%, o exhibe estas características en una combinación ponderada con tales y tales propiedades", podemos simplemente notar que si un sistema exhibe estos fenómenos en un grado suficiente entonces estará vivo, en virtud del significado del término. Si una relación de los hechos relevantes de bajo nivel determina los hechos acerca de la reproducción, la utilidad, el metabolismo, etc., de un sistema, entonces también determina los hechos acerca de si el sistema está *vivo*, en la medida que esta sea una cuestión fáctica.

Podemos sintetizar esto con un diagrama esquemático (fig. 2.1) que muestra cómo una propiedad de alto nivel P podría depender de dos parámetros A y B de bajo nivel, cada uno de los cuales puede tomar un rango de valores. Si tuviésemos una definición precisa en términos de condiciones necesarias y suficientes, entonces tendríamos algo así como la imagen de la izquierda, en la que el rectángulo oscuro representa la región en la que la propiedad P está instanciada. En cambio, la dependencia es invariablemente algo como la imagen de la derecha, en la que los límites son vagos y existe un área grande en la cual la cuestión de la existencia de la propiedad está indeterminada, pero también hay un área en la que la cuestión es clara. (Podría ser indeterminado que las bacterias o los virus informáticos estén vivos, pero no hay ninguna duda de que los perros lo están.) Dado un

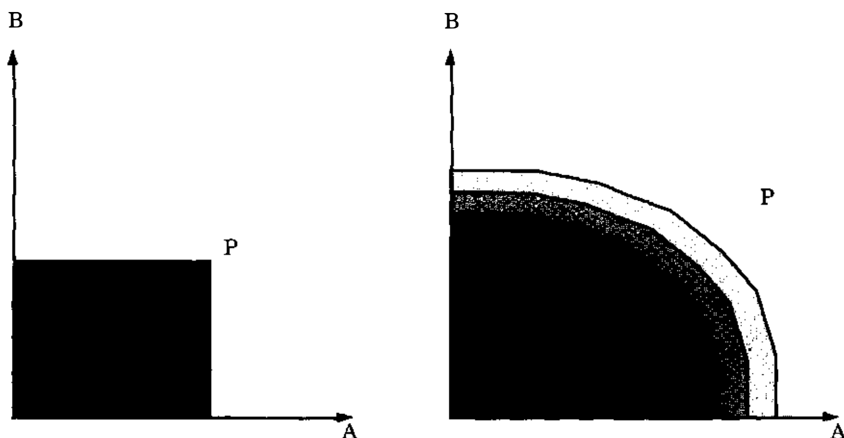



*Figura 2.1.* Dos modos en los que una propiedad P podría depender de las propiedades A y B.

ejemplo en el área definida, ejemplificar A y B en un grado suficiente como para que P lo esté, el condicional "Si $x$ es A y B en este grado, entonces $x$ es P" es una verdad conceptual, a pesar de la falta de una definición precisa de P. Cualquier indeterminación en dichos condicionales, en las áreas grises, reflejará una indeterminación en los hechos, que es como debería ser. La imagen puede extenderse directamente a la dependencia de una propiedad sobre un número arbitrario de factores y a los condicionales de superveniencia en general.

Es importante, entonces, que un conjunto de hechos puede *implicar* a otro conjunto sin que haya una definición precisa de las segundas nociones en términos de las primeras. El caso de más arriba proporciona un ejemplo: no existe ninguna definición simple de P en términos de A y B, pero los hechos acerca de A y B en una instancia implican los hechos acerca de P. Como otro ejemplo, piénsese acerca de la *redondez* de las curvas cerradas en el espacio bidimensional (fig. 2.2). Ciertamente, no existe ninguna definición perfecta de redondez en términos de nociones matemáticas más simples. Sin embargo, considérese la figura a la izquierda, especificada por la ecuación $2x^2 + 3y^2 = 1$. Existe un hecho verdadero —esta figura es redonda—, si es que hay hechos acerca de la redondez (compárese con la figura a la derecha, que ciertamente no es redonda). Además, este hecho está *implicado* por la descripción básica de la figura en términos matemáticos: dada la descripción y el concepto de redondez, el hecho de que la figura es redonda está determinado. Dado que los hechos A pueden implicar hechos B sin una definición de los hechos B en términos de los hechos A, la noción de superveniencia lógica no se ve afectada por la ausencia de definiciones. (Al pensar sobre

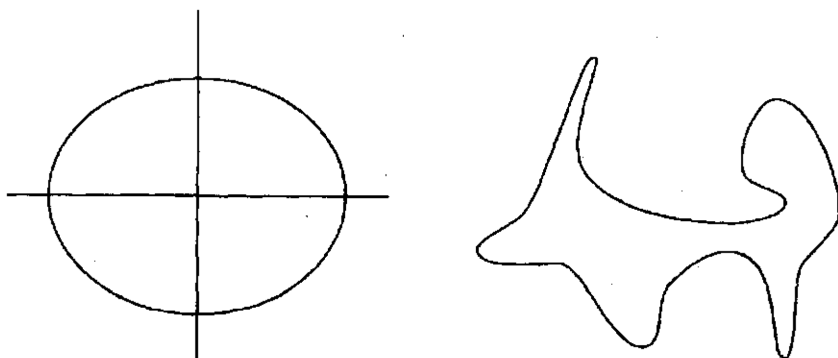


*Figura 2.2.* La curva redonda $2x^2 + 3y^2 = 1$ y una amiga no redonda.

cuestiones más complejas y sobre las objeciones a la superveniencia lógica, podría ser útil que tenga en mente este ejemplo.)

Podemos formular la cuestión diciendo que el tipo de "significado" de un concepto que es relevante en la mayoría de los casos no es una definición, sino una *intensión*: una función que especifica cómo el concepto se aplica a diferentes situaciones. A veces una intensión podría ser sintetizable en una definición, pero no es preciso que lo sea, como estos casos lo sugieren. Pero mientras haya un hecho acerca de cómo se aplican los conceptos en diversas situaciones, entonces tendremos una intensión; y como expondré en breve, ese es todo el "significado" que mis argumentos necesitarán.

## Revisabilidad

La segunda objeción, planteada por Quine (1951), es que las supuestas verdades conceptuales están siempre sujetas a revisión ante suficiente evidencia empírica. Por ejemplo, si la evidencia nos fuerza a revisar diversos enunciados generales en una teoría, es posible que un enunciado que una vez parecía ser conceptualmente verdadero pueda resultar falso.

Esto ocurre en el caso de muchas supuestas verdades conceptuales, pero no se aplica a los condicionales de superveniencia que estamos considerando, que tienen la forma "Si los hechos de bajo nivel resultan de un modo, entonces los hechos de alto nivel resultarán de otro". Los hechos especificados en el antecedente de este condicional incluyen todos los factores empíricos relevantes. La evidencia empírica podría mostrarnos que el antecedente del condicional es falso, pero no que el condicional es falso. En el caso extremo, podemos verificar que el antecedente constituya una especificación completa de los hechos de bajo nivel relacionados con el mundo. La propia amplitud del antecedente asegura que la evidencia empírica sea irrelevante para el valor de verdad del condicional. (Este cuadro se complica en cierta manera debido a la existencia de las necesidades *a posteriori*, que discutiré en breve. Aquí, sólo me ocupo de los condicionales epistémicos acerca de los modos como el mundo real podría resultar.)

Aunque las consideraciones acerca de la revisabilidad proporcionan un argumento plausible de que no hay muchas verdades conceptuales *breves*, nada en estas consideraciones va en contra del tipo de verdad conceptual restringida y compleja de la que me ocupo. La conclusión de estas observaciones es que las condiciones de verdad de un enunciado de alto nivel pueden no ser fácilmente *localizables*, ya que todo tipo de factores podría tener alguna clase de relevancia

indirecta; pero las condiciones de verdad globales proporcionadas por un condicional de superveniencia no están amenazadas. Si el significado define una función de mundos posibles en clases de referencia (una intensión) y si los mundos posibles son describibles de un modo finito (en términos de la disposición de cualidades básicas en esos mundos, digamos) entonces una vasta clase de condicionales conceptualmente verdaderos resultará en forma automática.

## La necesidad *a posteriori*

Tradicionalmente se consideró que todas las verdades conceptuales son cognoscibles *a priori*, al igual que las verdades necesarias, y que las tres clases de verdades —*a priori*, necesarias y conceptuales— están estrechamente relacionadas o son incluso coextensionales. El libro *Naming and Necessity* (1972) de Saul Kripke cuestionó esta concepción argumentando que existe una clase grande de enunciados necesariamente verdaderos cuya verdad no es cognoscible *a priori*. Un ejemplo es el enunciado "El agua es $H_2O$". No podemos saber *a priori* que esto es verdad; por todo lo que sabemos (o según todo lo que sabíamos al comienzo de la indagación), el agua podría estar hecha de alguna otra cosa, quizá XYZ. Kripke argumenta que, no obstante, dado que el agua es $H_2O$ en el mundo real, entonces es $H_2O$ en todos los mundos posibles. Se deduce entonces que "El agua es $H_2O$" es una verdad necesaria a pesar de su naturaleza *a posteriori*.

Esto plantea algunas dificultades para el marco teórico que he presentado. Por ejemplo, según algunas concepciones, estas verdades necesarias son verdades conceptuales, lo que implica que no todas las verdades conceptuales son cognoscibles *a priori*. Según otras concepciones alternativas, esos enunciados no son verdades conceptuales, pero entonces el vínculo entre la verdad conceptual y la necesidad se rompe. En diversos puntos de este libro, utilizaré métodos *a priori* para alcanzar alguna comprensión sobre la necesidad; este es el tipo de cosas que suele interpretarse que la concepción de Kripke puso en cuestión.

Si se lo analiza, pienso que puede advertirse que estas complicaciones no cambian nada fundamental en mis argumentos; pero vale la pena tomarse el trabajo de aclarar qué es lo que está ocurriendo. Emplearé algo de tiempo organizando un marco sistemático que permita ocuparse de estas cuestiones, que por otro lado aparecerán una y otra vez. En particular, presentaré un modo natural de capturar las tesis de Kripke en una imagen bidimensional del significado y la necesidad, Este marco es una síntesis de ideas sugeridas por Kripke, Putnam, Kaplan, Stalnaker, Lewis, Evans, Davies, Humberstone, y otros que estudiaron este fenómeno bidimensional.

Según el enfoque tradicional de la referencia, derivado de Frege aunque revestido aquí de la terminología moderna, un concepto define una función $f: W \rightarrow R$ de mundos posibles en referentes. Una función de este tipo suele denominarse una *intensión*; junto con una especificación de un mundo $w$, determina una *extensión* $f(w)$. Según el punto de vista de Frege, todo concepto tiene un *sentido*, que se supone determina la referencia del concepto según el estado del mundo; de modo que estos sentidos corresponden estrechamente a las intensiones. Solía pensarse el sentido como el *significado* del concepto en cuestión.

Trabajos más recientes reconocieron que ninguna única intensión puede hacer todo el trabajo que un significado debe hacer. El cuadro desarrollado por Kripke complica las cosas haciendo notar que la referencia en el mundo real y en los mundos posibles contrafácticos está determinada por mecanismos bastante diferentes. En cierto modo, puede interpretarse que la imagen kripkeana divide el cuadro fregeano en dos niveles separados.

El aporte de Kripke puede expresarse diciendo que existen, de hecho, *dos* intensiones asociadas a un concepto dado. Esto es, existen dos patrones bastante distintos en los que el referente de un concepto depende del estado del mundo. Primero, tenemos la dependencia por la cual se fija la referencia en el mundo *real* según cómo resulte el mundo: si resulta de un modo, un concepto seleccionará una cosa, pero si resulta de otro modo, el concepto seleccionará alguna otra cosa. Segundo, tenemos la dependencia según la cual se determina la referencia en los mundos *contrafácticos*, dado que la referencia en el mundo real ya ha sido fijada. En correspondencia con cada una de estas dependencias existe una intensión que llamaré intensiones *primaria* y *secundaria*, respectivamente.

La intensión *primaria* de un concepto es una función de mundos en extensiones que refleja el modo como se determina la referencia en el mundo real. En un mundo dado, selecciona cuál sería el referente del concepto si ese mundo resultase ser el real. Considérese el concepto "agua". Si el mundo real resultase tener XYZ en los océanos y lagos, entonces "agua" referiría a XYZ,[21] pero dado que resulta tener $H_2O$ en océanos y lagos, "agua" refiere al $H_2O$. De modo que la intensión primaria de "agua" aplica el mundo XYZ en XYZ, y el mundo $H_2O$ en $H_2O$. En una primera aproximación, podríamos decir que la intensión primaria selecciona el líquido claro y bebible dominante en los océanos y lagos; o más brevemente que selecciona la *sustancia acuosa* en un mundo.

Sin embargo, *dado* que "agua" refiere al $H_2O$ en el mundo real, Kripke hace notar (como lo hace Putnam, 1975) que es razonable decir que el agua es $H_2O$ en todo mundo contrafáctico. La *intensión*

*secundaria* de "agua" selecciona el agua en todo mundo contrafáctico; de modo que si Kripke y Putnam tienen razón, la intensión secundaria selecciona $H_2O$ en todos los mundos.[22]

Es la intensión primaria de un concepto la que es más importante para mis propósitos: para un concepto de un fenómeno natural, es la intensión primaria la que captura lo que debe explicarse. Si alguien dice "Explique el agua" mucho antes de que sepamos que el agua es de hecho $H_2O$, lo que buscamos es más o menos una explicación del líquido claro y bebible en su ambiente. Es sólo *después* de que la explicación se completa que sabemos que el agua es $H_2O$. La intensión primaria de un concepto, a diferencia de su intensión secundaria, es independiente de factores empíricos: la intensión *especifica* de qué manera la referencia depende del modo como resulta el mundo externo, de modo que ella misma no depende del modo como el mundo externo resulte.

Por supuesto, cualquier caracterización breve de la intensión primaria de un concepto del tipo "líquido claro y bebible dominante en el ambiente" será una simplificación. La verdadera intensión sólo puede determinarse a partir de la consideración detallada de escenarios específicos: ¿qué diríamos si el mundo resultase de esta forma? ¿Qué diríamos si el mundo resultase de aquella otra forma? Por ejemplo, si hubiese resultado que el líquido en los lagos es $H_2O$ pero el líquido en los océanos es XYZ, entonces probablemente habríamos dicho que ambos son agua; si la sustancia en los océanos y lagos fuese una mezcla de 95% A y 5% B, probablemente habríamos dicho que A pero no B es agua; si hubiese resultado que una sustancia que no es clara ni bebible tuviese una relación microfísica apropiada con el líquido claro y bebible en nuestro ambiente, probablemente también llamaríamos a esa sustancia "agua" (como lo hacemos en el caso del hielo o del "agua sucia"). Las condiciones completas de lo que se requiere para que algo califique como "agua" serán bastante vagas en la periferia y no tienen por qué ser inmediatamente evidentes cuando se piensa en ello, pero nada de esto hace mucha diferencia para la imagen que describo. Usaré "sustancia acuosa" como término de referencia para encapsular la intensión primaria, cualquiera sea esta.[23]

En ciertos casos, la decisión de cuál es la referencia de un concepto en el mundo real involucra una gran cantidad de reflexión acerca de qué es lo más razonable para decir; como, por ejemplo, con las preguntas acerca de la referencia de "masa" cuando el mundo real resultó ser uno en el cual la teoría de la relatividad general es verdadera,[24] o quizá con preguntas acerca de qué califica como "creencia" en el mundo real. De modo que la consideración de exacta-

mente qué es lo que la intensión primaria selecciona en diversos candidatos del mundo real puede suponer una cantidad correspondiente de reflexión. Pero, esto no significa que la cuestión no sea *a priori*: tenemos la capacidad de realizar este razonamiento independientemente de cómo resulte el mundo. Podría ocurrir que los informes de experimentos que confirman la relatividad sean cuestionables, de modo que no estamos seguros si el mundo actual resulta ser un mundo relativista: de cualquier modo, tenemos la capacidad para razonar acerca de cuál sería referencia del concepto "masa" *si* ese estado de cosas resultase ser el real.

(Surgen varias complicaciones cuando se analizan las intensiones primarias de los conceptos usados por los individuos dentro de una comunidad lingüística. Estas podrían manejarse notando que el concepto de un individuo puede tener una intensión primaria que supone una cierta deferencia con el concepto de la comunidad que lo rodea, de modo que mi concepto "olmo" podría seleccionar lo que aquellos en torno de mí llaman "olmos"; pero de cualquier manera este tipo de problema es irrelevante a las cuestiones de las que me ocuparé, para las cuales podemos suponer que hay sólo una persona en la comunidad, o que todos los individuos están igualmente bien informados o incluso que la comunidad es un individuo gigante. Podrían también surgir algunos problemas técnicos al usar intensiones primarias para construir una teoría semántica general, por ejemplo, ¿es la referencia de un concepto esencial para el concepto?, ¿podrían hablantes diferentes asociar distintas intensiones primarias con la misma palabra? Pero no intento construir aquí una teoría semántica completa, por lo que podemos evitar este tipo de preocupaciones.

A veces, los filósofos sospechan de entidades como las intensiones primarias porque ven en ellas una reminiscencia de una teoría "descriptiva" de la referencia. Pero las descripciones no tienen ningún papel esencial en este marco; las utilizo meramente para detallar algunas de las características de las funciones relevantes de los mundos posibles en las extensiones. Es la función en sí misma, más que cualquier descripción sintetizadora, lo que es verdaderamente fundamental. Esta imagen es bastante compatible con la teoría "causal" de la referencia: simplemente debemos notar que la intensión primaria de un concepto como "agua" puede requerir una conexión causal apropiada entre el referente y el sujeto. Nos vemos llevados a creer en una teoría causal de la referencia, en primer lugar, precisamente debido a que consideramos los diversos modos en los que el mundo real podría resultar, y advertimos cuál resultaría ser el referente del concepto en esos casos; esto es, evaluamos la intensión primaria de un concepto en esos mundos.)

Dado que la referencia de "agua" en el mundo real se fija seleccionando la sustancia acuosa, podríamos pensar que el agua es sustancia acuosa en todos los mundos posibles. Kripke y Putnam señalaron que esto no es así: si el agua es $H_2O$ en el mundo real, entonces el agua es $H_2O$ en todos los mundos posibles. En un mundo (la "Tierra gemela" de Putnam) en el que el líquido claro y bebible dominante es XYZ y no $H_2O$, este líquido no es agua; es meramente sustancia acuosa. Todo esto es capturado por la intensión *secundaria* de "agua", que selecciona el agua en todos los mundos: esto es, selecciona $H_2O$ en todos los mundos.

La intensión secundaria de un concepto como "agua" no está determinado *a priori*, ya que depende de cómo las cosas resultan en el mundo concreto. Pero conserva una relación estrecha con la intensión primaria de más arriba. En este caso, la intensión secundaria se determina primero evaluando la intensión primaria en el mundo real, y luego *rigidificando* esta evaluación de modo que el mismo tipo de cosas se seleccione en todos los mundos posibles. Dado que la intensión primaria ("sustancia acuosa" selecciona al $H_2O$ en el mundo actual, se deduce de la rigidificación que la intensión secundaria selecciona al $H_2O$ en todos los mundos posibles.

Podemos sintetizar esto diciendo que "agua" es conceptualmente equivalente a "*dthat* (sustancia acuosa)", donde *dthat* es una versión del operador de rigidificación de Kaplan que convierte una intensión en un designador rígido por evaluación en el mundo real (Kaplan, 1979). La intensión fregeana única ha sido fragmentada en dos: una intensión primaria ("sustancia acuosa") que fija la referencia en el mundo real, y una intensión secundaria ("$H_2O$") que selecciona la referencia en los mundos contrafácticos posibles y que depende de cómo resulte el mundo real.

(Existe a veces una tendencia a suponer que una necesidad *a posteriori* hace que el análisis conceptual *a priori* sea irrelevante, pero esta suposición es infundada. Antes de que tan siquiera lleguemos al punto en que la designación rígida o similares se vuelvan relevantes, tenemos una historia que contar acerca de qué hace que un X del mundo actual *califique* como el referente de "X". Esta historia sólo puede contarse mediante un análisis de la intensión primaria. Y este proyecto es una empresa *a priori*, ya que involucra cuestiones acerca de a qué *referiría* nuestro concepto si el mundo real resultase de diversos modos. Dado que tenemos la capacidad de conocer a qué refieren nuestros conceptos cuando sabemos cómo resulta el mundo, entonces tenemos la capacidad de conocer a qué referirían nuestros conceptos *si* el mundo real resultase de diversos modos.Que el mundo real *resulte* o no de cierto modo hace poca

diferencia para responder esta cuestión, excepto que concentra nuestra atención.)

Tanto la intensión primaria como la secundaria pueden considerarse funciones *f*: $W \to R$ de mundos posibles en extensiones, donde los mundos posibles en cuestión se consideran de modos sutilmente diferentes. Podríamos decir que la intensión primaria selecciona el referente de un concepto en un mundo cuando se lo *considera el real* —esto es, cuando se lo considera un candidato del mundo real del pensador— mientras que la intensión secundaria selecciona el referente de un concepto en un mundo cuando se lo *considera un contrafáctico*, dado que el mundo real del pensador ya está fijado. Cuando se considera real al mundo XYZ, mi término "agua" selecciona XYZ en el mundo, pero cuando se lo considera contrafáctico, "agua" selecciona el $H_2O$.

La distinción entre estos dos modos de ver el mundo corresponde estrechamente a la distinción de Kaplan (1989) entre el *contexto de emisión* de una expresión y las *circunstancias de evaluación*. Cuando consideramos un mundo *w* como contrafáctico, mantenemos el mundo actual como el contexto de la emisión, pero utilizamos a *w* como una circunstancia de evaluación. Por ejemplo, si enuncio "Hay agua en el océano" en este mundo y la *evalúo* en el mundo XYZ, "agua" refiere al $H_2O$ y el enunciado es falso. Pero cuando consideramos a *w* como real, lo pensamos como un contexto potencial de emisión y nos preguntamos cómo serían las cosas si el contexto de la expresión resultase ser *w*. Si el contexto de mi oración "Hay agua en el océano" resultase ser el mundo XYZ, entonces el enunciado sería verdadero cuando se evalúa en ese mundo. La intensión primaria está por lo tanto estrechamente relacionada con lo que Kaplan llama el *carácter* de un término, aunque hay unas pocas diferencias,[25] y la intensión secundaria corresponde a lo que llama el *contenido* de un término.

Existe una leve asimetría en el sentido de que un contexto de emisión pero no la circunstancia de evaluación es lo que Quine (1969) llama un mundo posible *centrado*. Este es un par ordenado consistente de un mundo y un *centro* que representa el punto de vista dentro de ese mundo de un agente que utiliza el término en cuestión: el centro consiste en (al menos) un individuo y tiempo "marcados". (Esta sugerencia proviene de Lewis, 1979; Quine sugiere que el centro podría ser un punto en el espaciotiempo.) Un centro de esta clase es necesario para capturar el hecho de que un término como "agua" selecciona una extensión diferente para mí que para mi gemelo en Tierra Gemela, a pesar del hecho de que vivimos en el mismo universo.[26] Sólo nuestra posición en el universo difiere, y es esta

posición lo que constituye una diferencia relevante para el proceso de fijar la referencia.

Este fenómeno surge de un modo especialmente obvio para las términos indicadores como "yo", cuya referencia depende claramente de quién está utilizando el término y no sólo del estado global del mundo: la intensión primaria de "yo" selecciona al individuo en el centro de un mundo centrado. (La intensión secundaria de mi concepto "yo" selecciona a David Chalmers en todos los mundos posibles.) Existe un elemento indicador menos explícito en nociones como "agua", sin embargo, que puede analizarse aproximadamente como "*dthat* (el líquido claro y bebible predominante *en nuestro ambiente*)".[27] Es este elemento indicador el que requiere que las intensiones primarias dependan de mundos centrados. Una vez que la referencia al mundo real ha sido fijada, sin embargo, no se necesita ningún centro para evaluar la referencia en un mundo contrafáctico. La circunstancia de evaluación puede, por lo tanto, representarse mediante un mundo posible simple sin un centro.

Todo esto puede formalizarse notando que la historia completa de la referencia en los mundos contrafácticos no está determinada *a priori* por una función de un solo parámetro $f: W \rightarrow R$. En cambio, la referencia en un mundo contrafáctico depende de ese mundo y del modo como el mundo real resulta. Esto es, un concepto determina una función de dos parámetros

$$F: W^* \times W \rightarrow R$$

donde $W^*$ es el espacio de mundos posibles centrados, y $W$ es el espacio de mundos posibles ordinarios. El primer parámetro representa los contextos de emisión, o modos en los que podría resultar el mundo real, mientras que el segundo parámetro representa circunstancias de evaluación, o mundos contrafácticos posibles. De modo equivalente, un concepto determina una familia de funciones

$$F_v: W \rightarrow R$$

para cada $v \in W^*$ que representa un modo como el mundo real podría resultar, donde $F_v(w) = F(v, w)$. Para "agua", si $a$ es un mundo en el cual la sustancia acuosa es $H_2O$, entonces $F_a$ selecciona al $H_2O$ en cualquier mundo posible. Dado que en nuestro mundo el agua resultó ser $H_2O$, esta $F_a$ especifica las condiciones correctas de aplicación de "agua" a través de mundos contrafácticos. Si nuestro mundo hubiese resultado ser un mundo diferente $b$ en el cual la

sustancia acuosa era XYZ, entonces las condiciones de aplicación relevantes habrían sido especificadas por $F_b$, una intensión diferente que selecciona a XYZ en cualquier mundo posible.

La función $F$ se determina *a priori*, ya que todos los factores *a posteriori* están incluidos en sus parámetros. A partir de $F$ podemos recuperar nuestras dos intensiones de un solo parámetro. La intensión primaria es la función $f$: $W^* \rightarrow R$ determinada por la aplicación "diagonal" $f$: $w \mapsto F(w, w')$, donde $w'$ es idéntico a $w$ excepto que se ha eliminado el centro. Esta es la función mediante la cual se fija la referencia en el mundo real. La intensión secundaria es la aplicación $F_a$: $w \mapsto F(a, w)$, donde $a$ es nuestro mundo real. Esta intensión selecciona la referencia en mundos contrafácticos. Una consecuencia inmediata es que la intensión primaria y la intensión secundaria coinciden en su aplicación al mundo real: $f(a) = F_a(a') = F(a, a')$.

En la dirección inversa, la función de dos parámetros $F$ y por lo tanto la intensión secundaria $F_a$ pueden derivarse usualmente de la intensión primaria $f$, con la ayuda de una "regla" acerca de cómo la intensión secundaria depende de la intensión primaria y del mundo real $a$. Esta regla depende del tipo de concepto. Para un concepto que es un designador rígido, la regla es que en un mundo $w$, la intensión secundaria selecciona en $w$ lo que la intensión primaria selecciona en $a$ (o quizá, para los términos de tipos naturales, lo que tenga la misma estructura subyacente de lo que la intensión primaria selecciona en $a$). Más formalmente, sea $D$: $R \times W \rightarrow R$ un operador de "proyección" que va desde una clase seleccionada en algún mundo a miembros de "esa" clase en otro mundo posible. Entonces la intensión secundaria Fa es exactamente la función $D(f(a),-)$, que podemos pensar como *dthat* aplicada a la intensión dada por $f$.

Para otros conceptos, la derivación de la intensión secundaria a partir de la intensión primaria será más fácil. Para expresiones "descriptivas" como "doctor", "cuadrado" y "sustancia acuosa", la designación rígida no desempeña ningún papel especial: se aplican a mundos contrafácticos independientemente de cómo resulte ser el mundo actual. En estos casos, la intensión secundaria es una copia simple de la intensión primaria (excepto por las diferencias debidas al centrado). El marco que esbocé puede manejar ambos tipos de conceptos.

Los términos de *propiedades*, como "caliente", pueden representarse de dos maneras en un marco intensional. Podemos considerar la intensión de una propiedad como una función de un mundo en una clase de individuos (los individuos que instancian la propiedad), o de un mundo en las propiedades mismas. Cualquiera de los dos

modos es compatible con el marco actual; podemos fácilmente encontrar una intensión primaria y una intensión secundaria en los dos casos, y es fácil moverse entre los dos enfoques. Por lo general haré las cosas del primer modo, sin embargo, así como la intensión primaria de "caliente" seleccionará las entidades que califican como "calientes" en el mundo real, según cómo este resulte, la intensión secundaria seleccionará las cosas calientes en un mundo contrafáctico, dado que el mundo real resultó como lo hizo.

Las intensiones primaria y secundaria pueden pensarse como candidatas del "significado" de un concepto. Creo que no tiene sentido elegir una de estas como *el* significado; el término "significado" aquí es fundamentalmente honorífico. Podemos también pensar las intensiones primaria y secundaria como los aspectos *a priori* y *a posteriori* del significado, respectivamente.

Si hacemos esta equiparación, las dos intensiones respaldarán un cierto tipo de verdad conceptual, o verdad en virtud del significado. La intensión primaria respalda las verdades *a priori*, tal como "El agua es sustancia acuosa". Un enunciado de este tipo será verdadero independientemente de cómo resulte el mundo, aunque podría no valer en todos los mundos posibles no reales. La intensión secundaria no respalda las verdades *a priori*, pero sí las verdades válidas en todos los mundos contrafácticos posibles, tal como "El agua es $H_2O$". Ambos tipos califican como verdades en virtud del significado; simplemente son verdaderas en virtud de diferentes aspectos del significado.

También es posible considerarlas dos variedades de la verdad *necesaria*. La segunda corresponde a la interpretación más estándar de una verdad necesaria. La primera, sin embargo, puede también interpretarse como verdad a través de mundos posibles, siempre que esos mundos posibles se interpreten como contextos de emisión, o como modos en los que el mundo real podría resultar. Según esta interpretación sutilmente diferente, un enunciado S es necesariamente verdadero si, independientemente de cómo resulte el mundo real, S es verdadero. Si el mundo real resulta ser un mundo en el cual la sustancia acuosa es XYZ, entonces mi enunciado "XYZ es agua" será verdadero. Así, de acuerdo con esta interpretación de cuáles mundos posibles son *considerados reales*, "El agua es sustancia acuosa" es una verdad necesaria.

Esta clase de necesidad es lo que Evans (1979) llama "necesidad profunda", en oposición a las necesidades "superficiales" como "El agua es $H_2O$". Davies y Humberstone (1980) la analizan en detalle mediante un operador modal que llaman "fijamente real". La necesi-

dad profunda, a diferencia de la necesidad superficial, no está afectada por consideraciones *a posteriori*. Estas dos variedades de posibilidad y necesidad se aplican siempre a *enunciados*. Sólo hay un tipo relevante de posibilidad de *mundos*; los dos enfoques difieren en cómo se evalúa la verdad de un enunciado en un mundo.

Podemos ver esto de un modo diferente notando que existen dos conjuntos de *condiciones de verdad* asociados a cualquier enunciado. Si evaluamos los términos en un enunciado de acuerdo con sus intensiones primarias, llegamos a las condiciones de verdad *primarias* del enunciado; esto es, un conjunto de mundos posibles centrados en los cuales el enunciado, evaluado de acuerdo con las intensiones primarias de los términos allí contenidos, resulta ser verdadero. Las condiciones de verdad primarias nos cuentan cómo debe ser el mundo real para que una emisión del enunciado sea verdadera en ese mundo; esto es, especifican aquellos *contextos* en los cuales el enunciado resultaría ser verdadero. Por ejemplo, las condiciones de verdad primarias de "El agua está mojada" especifican aproximadamente que una emisión de este tipo será verdadera en el conjunto de mundos en los que la sustancia acuosa está mojada.

Si en cambio evaluamos los términos involucrados según sus intensiones secundarias, llegamos a las más familiares *condiciones de verdad secundarias*. Estas condiciones especifican el valor de verdad de un enunciado en mundos contrafácticos, dado que el mundo actual resultó como lo hizo. Por ejemplo, las condiciones de verdad secundarias de "El agua está mojada" (emitida en este mundo) especifica aquellos mundos en los cuales el $H_2O$ está mojado. Nótese que no hay ningún peligro de ambigüedad en la verdad del mundo real; las condiciones de verdad primarias y secundarias siempre especificarán el mismo valor de verdad cuando se las evalúa en el mundo real.

Si consideramos una proposición como una función de mundos posibles en valores de verdad, entonces estos dos conjuntos de condiciones de verdad producen dos *proposiciones* asociadas a cualquier enunciado. La composición de las intensiones primarias de los términos involucrados produce una *proposición primaria*, que es válida precisamente en esos contextos de emisión en los que resultaría que el enunciado expresa una verdad. (Esta es la "proposición diagonal" de Stalnaker, 1978. Estrictamente hablando, es una proposición centrada, o una función de mundos centrados en valores de verdad.) Las intensiones secundarias producen una proposición *secundaria*, que es válida en aquellas circunstancias contrafácticas en las que el enunciado, tal como se lo emite en el mundo real, es verdadero. La proposición secundaria es el "contenido" de una emi-

sión según Kaplan y más comúnmente se la interpreta como la proposición expresada por un enunciado, pero la proposición primaria también es fundamental.

Los dos tipos de verdad necesarias de un enunciado corresponden precisamente a la necesidad de los dos tipos de proposiciones asociadas. Un enunciado es necesariamente verdadero en el primer sentido (*a priori*) si la proposición primaria asociada es válida en todos los mundos centrados posibles (esto es, si resulta que el enunciado expresa una verdad en cualquier contexto de emisión). Un enunciado es necesariamente verdadero en el sentido *a posteriori* si la proposición secundaria asociada es válida en todos los mundos posibles (esto es, si la sentencia tal como es emitida en el mundo *real* es verdadera en todos los mundos contrafácticos). La primera corresponde a la necesidad profunda de Evans y la segunda a la más familiar necesidad superficial.

Para ilustrar, considérese el enunciado "El agua es $H_2O$". Las intensiones primarias de "agua" y "$H_2O$" difieren, de modo que no podemos saber *a priori* que el agua sea $H_2O$; la proposición *primaria* asociada no es necesaria (es válida en aquellos mundos centrados en los que la sustancia acuosa tiene una cierta estructura molecular). Sin embargo, las intensiones secundarias coinciden, de modo que "El agua es $H_2O$" es verdadera en todos los mundos posibles cuando se los evalúa de acuerdo con las intensiones *secundarias*, esto es, la proposición *secundaria* asociada es necesaria. La necesidad *a posteriori* de Kripke surge justamente cuando las intensiones secundarias en un enunciado respaldan una proposición necesaria, pero las intensiones primarias no.

Considérese en cambio el enunciado "El agua es sustancia acuosa". Aquí las intensiones primarias asociadas a "agua" y "sustancia acuosa" son una misma, de modo que podemos saber que este enunciado es verdadero *a priori*, en tanto poseamos los conceptos. La proposición primaria asociada es necesaria, de modo que este enunciado es necesariamente verdadero en el sentido "profundo" de Evans. Sin embargo, las intensiones secundarias difieren, ya que "agua" está rigidificada, pero "sustancia acuosa" no: en un mundo en el que XYZ es el líquido claro y bebible, la intensión secundaria de "sustancia acuosa" selecciona a XYZ pero la intensión de "agua" no. Por lo tanto, la proposición *secundaria* asociada no es necesaria, y el enunciado no es una verdad necesaria en el sentido más familiar; es un ejemplo del "*a priori* contingente" de Kripke.

En general, muchos aparentes "problemas" que surgen de estas consideraciones kripkeanas son una consecuencia de intentar comprimir la imagen doblemente parametrizada de la referencia en una

sola noción de significado o de necesidad. Por lo general, estos problemas pueden eliminarse notando explícitamente el carácter bidimensional de la referencia y distinguiendo cuidadosamente la noción de significado o de necesidad que esté en juego.[28]

También es posible utilizar este marco bidimensional para formular una concepción de la semántica del *pensamiento*, así como del lenguaje. Hago esto con mucha mayor extensión en otro lugar (Chalmers, 1994c). Este aspecto de nuestro marco teórico no será fundamental aquí, pero vale la pena mencionarlo, ya que surgirá en uno o dos lugares menores. La idea básica es muy similar: dado un *concepto* en el pensamiento de un individuo, podemos asignarle una intensión primaria que corresponde a lo que seleccionará según cómo resulte el mundo real y una intensión secundaria que corresponde a lo que selecciona en mundos contrafácticos, dado que el mundo real resulta como es. De forma similar, dada una *creencia*, podemos asignarle una proposición primaria y una proposición secundaria (lo que en otro lugar llamo el contenido "nocional" y "relacional" de la creencia).

Por ejemplo, conceptos como "Héspero" y "Fósforo" tendrán diferentes intensiones primarias (uno selecciona la estrella vespertina en un mundo centrado determinado, el otro selecciona la estrella matutina), pero la misma intensión secundaria (ambas seleccionan a Venus en todos los mundos). El pensamiento "Héspero es Fósforo" tendrá una proposición primaria verdadera en todos los mundos centrados en los cuales la estrella vespertina es la estrella matutina: el hecho de que este pensamiento es informativo en lugar de trivial corresponde al hecho de que la proposición primaria es contingente, ya que las intensiones primarias de los dos términos difieren.

La proposición primaria, más que la proposición secundaria, captura la apariencia de las cosas desde el punto de vista del sujeto: produce el conjunto de mundos centrados que el sujeto, por tener la creencia, avala como ambientes potenciales en los que podría estar viviendo (al creer que Héspero es Fósforo, yo avalo todos aquellos mundos centrados en los cuales la estrella vespertina y la estrella matutina en torno del centro son idénticas). Es también bastante fácil argumentar que la proposición primaria, y no la proposición secundaria, gobierna las relaciones cognitivas y racionales entre los pensamientos. Por esta razón es natural pensar en la proposición primaria como el contenido *cognitivo* de un pensamiento.[29]

## Necesidad lógica, verdad conceptual y conceptibilidad

Habiendo delineado nuestro marco teórico, podemos ahora especificar las relaciones entre la necesidad lógica, la verdad conceptual y la conceptibilidad. Comenzando por la necesidad lógica: esta es sólo necesidad como se explicó antes. Un enunciado es lógicamente necesario si y sólo si es verdadero en todos los mundos lógicamente posibles. Por supuesto, tenemos dos variedades de necesidad lógica de enunciados, según que evaluemos la verdad en un mundo posible de acuerdo con las intensiones primarias o secundarias. Podríamos llamar a estas variedades *necesidad 1* y *necesidad 2*, respectivamente.

Este análisis explicita la necesidad y posibilidad lógica de un enunciado en términos de a) la posibilidad lógica de *mundos*, y b) las intensiones determinadas por los términos involucrados en el enunciado. Ya analizamos las intensiones. En lo que concierne a la noción de mundo lógicamente posible, esta es una especie de primitiva: como antes, podemos de una manera intuitiva pensar en un mundo lógicamente posible como un mundo que Dios podría haber creado (dejando de lado las preguntas acerca del propio Dios). No me involucraré en la fastidiosa cuestión de la condición ontológica de estos mundos, sino que simplemente los aceptaré como una herramienta, del mismo modo como aceptamos la matemática.[30] En lo que concierne a la *extensión* de la clase, la característica más importante es que todo mundo concebible es lógicamente posible, una cuestión de la que tendré más que decir en un momento.

En lo que concierne a la verdad conceptual, si equiparamos el significado con la intensión (primaria o secundaria), es fácil hacer el vínculo entre la verdad en virtud del significado y la necesidad lógica. Si un enunciado es lógicamente necesario, su verdad será un subproducto automático de las intensiones de los términos (y la estructura composicional del enunciado). No necesitamos introducir en el mundo ningún otro papel, ya que las intensiones en cuestión serán satisfechas en todo mundo posible. De modo similar, si un enunciado es verdadero en virtud de sus intensiones, será verdadero en todo mundo posible.

Como antes, hay dos variedades de verdades conceptuales, según que equiparemos los "significados" con las intensiones primarias o secundarias, haciendo un paralelo con las dos variedades de verdad necesarias. Si tomamos decisiones paralelas en los dos casos, un enunciado es conceptualmente verdadero si y sólo si es necesariamente verdadero. El enunciado "El agua es sustancia acuosa" es conceptualmente verdadero y necesariamente verdadero en el primer

sentido; y "El agua es $H_2O$" es conceptualmente verdadero y necesariamente verdadero en el segundo. Sólo la primera variedad de verdad conceptual será en general accesible *a priori*. La segunda variedad incluirá muchas verdades *a posteriori*, ya que la intensión secundaria depende de cómo resulte el mundo real.

(No sostengo que las intensiones sean *el* modo correcto de pensar sobre los significados. El significado es una noción multifacética, y algunas de sus facetas podrían no estar reflejadas perfectamente por las intensiones, de modo que podríamos resistirnos a la equiparación de los dos, al menos en algunos casos.[31] Más bien, debería considerarse que la equiparación del significado y la intensión es estipulativa: si hacemos la equiparación, entonces podemos hacer varias conexiones útiles. No mucho depende del uso de la palabra "significado". En cualquier caso, la verdad en virtud de la intensión es el único tipo de verdad en virtud del significado que necesitaremos.)

También podemos establecer un vínculo entre la posibilidad lógica de los enunciados y su *conceptibilidad*, si somos cuidadosos. Digamos que un enunciado es concebible (o concebiblemente verdadero) si es verdadero en todos los mundos concebibles. Esto no debe confundirse con otros sentidos de "concebible". Por ejemplo, existe un sentido según el cual un enunciado es concebible si según todo lo que sabemos es verdadero, o si no sabemos que sea imposible. En este sentido, la conjetura de Goldbach y su negación son concebibles. Pero el miembro falso del par no calificará como concebible en el sentido en el que estoy usando el término, ya que no hay ningún mundo concebible en el que sea verdadero (es falso en todos los mundos).

Según este enfoque de la noción, la conceptibilidad de un enunciado involucra dos cosas: primero, la conceptibilidad de un mundo relevante, y segundo, la verdad del enunciado en ese mundo.[32] Se deduce que al hacer juicios de conceptibilidad, debemos asegurarnos de que describimos el mundo que estamos concibiendo de una forma correcta, evaluando apropiadamente la verdad de un enunciado en el mundo. Podríamos pensar a primera vista que es concebible que la conjetura de Goldbach sea falsa, concibiendo un mundo en el que los matemáticos anuncian que lo es; pero si de hecho la conjetura de Goldbach fuera verdadera, entonces estaríamos *describiendo erróneamente* este mundo; es en realidad un mundo en el que la conjetura es verdadera y algunos matemáticos cometieron un error.

En la práctica, para hacer un juicio de conceptibilidad, sólo necesitamos considerar una *situación* concebible —una pequeña parte de un mundo— y luego asegurarnos de que la estamos describiendo correctamente. Si hay una situación concebible en la que un enunciado es verdadero, obviamente habrá un mundo concebible en

el cual el enunciado es verdadero, de modo que este método producirá resultados razonables a la vez que significará una carga menor sobre nuestros recursos cognitivos que concebir un mundo completo.

A veces se dice que ejemplos como "El agua es XYZ" muestran que la conceptibilidad no implica la posibilidad, pero creo que la situación es más sutil que esto. En efecto, existen dos variedades de conceptibilidad, que podríamos llamar *conceptibilidad-1* y *conceptibilidad-2,* según que evaluemos un enunciado en un mundo concebible de acuerdo con las intensiones primarias o secundarias de los términos involucrados. "El agua es XYZ" es concebible-1, ya que existe un mundo concebible en el cual el enunciado (evaluado de acuerdo con las intensiones primarias) es verdadero, pero no es concebible-2, ya que no existe ningún mundo concebible en el cual el enunciado (evaluado de acuerdo con la intensión secundaria) sea verdadero. Estos dos tipos de conceptibilidad reflejan precisamente los dos tipos de posibilidad lógica mencionados antes.

Suele equipararse la conceptibilidad de un enunciado a la conceptibilidad-1 (el sentido en el cual "El agua es XYZ" es concebible), ya que es este tipo de conceptibilidad el que es accesible *a priori.* Y con mayor frecuencia todavía, suele equipararse la *posibilidad* de un enunciado a la posibilidad 2 (el sentido en el cual "El agua es XYZ" es imposible). Interpretada de *este* modo, la conceptibilidad no implica posibilidad. Pero sigue siendo el caso que la conceptibilidad-1 implica la posibilidad-1, y la conceptibilidad-2 implica la posibilidad-2. Simplemente debemos tener cuidado de no juzgar la conceptibilidad-1 cuando lo relevante es la posibilidad-2. Esto es, debemos ser cuidadosos de no describir el mundo que estamos concibiendo (el mundo XYZ, digamos) según las intensio-nes primarias, cuando sería más apropiado utilizar las intensiones secundarias.[33]

Se deduce de todo esto que la distinción citada frecuentemente entre la posibilidad "lógica" y la posibilidad "metafísica" que se origina en los casos de Kripke —en los que se sostiene que es lógicamente posible, pero no metafísicamente posible, que el agua sea XYZ— no es una distinción en el nivel de los *mundos*, sino, cuanto más, una distinción en el nivel de los *enunciados*. Un enunciado es "lógicamente posible" en este sentido si es verdadero en algún mundo cuando se lo evalúa de acuerdo con las intensiones primarias; un enunciado es "metafísicamente posible" si es verdadero en algún mundo cuando se lo evalúa de acuerdo con las intensiones secundarias. El espacio relevante de mundos es el mismo en ambos casos.[34]

Muy importante, ninguno de los casos que hemos considerado nos da razones para pensar que algún *mundo* concebible sea imposible. Las preocupaciones acerca de la distancia entre la conceptibi-

lidad y la posibilidad se aplican al nivel de los enunciados, no a los mundos: o usamos un enunciado para describir erróneamente un mundo concebido (como en el caso de Kripke, y el segundo caso de Goldbach), o sostenemos que un enunciado es concebible sin concebir un mundo para nada (como en el primer caso de Goldbach). De modo que no parece haber ninguna razón para negar que la conceptibilidad de un mundo implique su posibilidad. De aquí en más daré esto por sentado como una aseveración acerca de la posibilidad lógica; cualquier variedad de la posibilidad para la que la conceptibilidad no implique la posibilidad será, entonces, una clase más restringida. Alguien podría sostener que existe una variedad más estrecha de "mundos metafísicamente posibles", pero cualquier razón para creer en una clase de este tipo debería ser independiente de las razones estándar que consideré aquí. De cualquier manera, es la posibilidad lógica la que es fundamental en las cuestiones acerca de la explicación. (Una modalidad "metafísica" más fuerte podría, en el mejor de los casos, ser relevante para cuestiones acerca de ontología, materialismo, etc.; la discutiremos cuando esas cuestiones se vuelvan relevantes en el capítulo 4.)

Una implicación en la dirección opuesta, de la posibilidad lógica a la conceptibilidad, es más truculenta, en el sentido de que los límites de nuestra capacidad cognitiva implican que existen algunas situaciones posibles que no podemos concebir, quizá debido a su mayor complejidad. Sin embargo, si entendemos la conceptibilidad como conceptibilidad en principio —quizá conceptibilidad por un superser— entonces es plausible que la posibilidad lógica de un mundo implique su conceptibilidad y, por lo tanto, que la posibilidad lógica de un enunciado implique su conceptibilidad (en el sentido relevante). De cualquier modo, me ocuparé más bien de la otra implicación.

Si un enunciado es lógicamente posible o necesario de acuerdo con su intensión primaria, la posibilidad o necesidad es cognoscible *a priori*, al menos en principio. La modalidad no es epistémicamente inaccesible: la posibilidad de un enunciado es una función de las intensiones involucradas y el espacio de mundos posibles, donde ambos son epistémicamente accesibles en principio, y ninguno de los cuales depende de hechos *a posteriori* en este caso. De esta manera, las cuestiones sobre la posibilidad-1 y la conceptibilidad-1 son en principio accesibles desde un sillón. En cambio, las cuestiones de la posibilidad-2 y la conceptibilidad-2 sólo serán, en muchos casos, accesibles *a posteriori*, ya que los hechos acerca del mundo externo pueden tener un papel en la determinación de las intensiones secundarias.

La clase de las verdades necesarias 1 corresponde directamente a la clase de las verdades *a priori*. Si un enunciado es verdadero *a*

*priori*, entonces es verdadero independientemente de cómo resulte el mundo actual; esto es, es verdadero en todos los mundos considerados reales, de modo que es necesario 1. Y recíprocamente, si un enunciado es necesario-1, entonces será verdadero independientemente de cómo resulte el mundo actual, de modo que será verdadero *a priori*. En la mayoría de este tipo de casos, la verdad del enunciado será cognoscible por nosotros *a priori*; podrían ser excepciones ciertos enunciados matemáticos cuya verdad no podemos determinar y ciertos enunciados que son tan complejos que no podemos comprenderlos. Aun en estos casos, parece razonable decir que son cognoscibles *a priori* al menos *en principio*, aunque estén más allá de nuestra limitada capacidad cognitiva. (Volveré a la cuestión cuando esta sea relevante más adelante.)

## Necesidad lógica y superveniencia lógica

Obtenemos dos nociones levemente diferentes de superveniencia lógica si utilizamos las clases primaria o secundaria de necesidad lógica. Si "glup" tiene asociadas una intensión primaria y una secundaria, entonces la "glupidad" puede supervenir lógicamente a las propiedades físicas de acuerdo con la intensión primaria o secundaria de "glup". La superveniencia de acuerdo con la intensión secundaria —esto es, superveniencia con necesidad *a posteriori* como la modalidad relevante— corresponde a lo que algunos llaman "superveniencia física", pero acabamos de ver que puede considerarse una variedad de superveniencia lógica.

(En realidad, sólo hay un tipo de superveniencia lógica de *propiedades*, así como sólo hay un tipo de necesidad lógica de las proposiciones. Pero, hemos visto que los términos o conceptos determinan efectivamente dos propiedades, una por medio de una intensión primaria ["sustancia acuosa"] y la otra por medio de una intensión secundaria ["$H_2O$"]. De modo que para un concepto dado ["agua"], existen dos modos en los que las propiedades asociadas a ese concepto podrían supervenir. A veces hablaré informalmente de las intensiones primaria y secundaria asociadas a una propiedad, y de los dos modos en los que una propiedad podría supervenir.)

Analizaré las versiones primaria y secundaria de la superveniencia en casos específicos, pero la primera será más importante. Especialmente cuando se consideran cuestiones acerca de la explicación, las intensiones primarias son más importantes que las intensiones secundarias. Como se hizo notar antes, sólo podemos trabajar con la intensión primaria al comienzo de una indagación y es esta intensión la que determina si una explicación es o no satisfactoria. Para

explicar el agua, por ejemplo, tenemos que explicar cosas como su claridad, liquidez, etc. La intensión secundaria ("$H_2O$") no surge sino hasta después de que se completó la explicación y, por lo tanto, no determina un criterio del éxito explicativo. Es la superveniencia lógica de acuerdo con la intensión primaria lo que determina si una explicación reductiva es posible. Si no especifico otra cosa, es la superveniencia lógica de acuerdo con la intensión primaria lo que por lo general será el motivo de discusión.

Si elegimos un tipo de intensión —digamos, la intensión primaria— y nos apegamos a ella, entonces podemos ver que diversos modos de formular la superveniencia lógica son equivalentes. De acuerdo con la definición dada al comienzo de este capítulo, las propiedades B son lógicamente supervenientes a la propiedades A si para cualquier situación lógicamente posible *Y* que es A indiscernible de una situación real *X*, entonces todos los hechos B verdaderos en *X* son verdaderos en *Y*. O más simplemente, las propiedades B son lógicamente supervenientes a las propiedades A si para cualquier situación real *X*, los hechos A acerca de *X implican* los hechos B acerca de *X* (donde "P implica Q" se entiende como "Es lógicamente imposible que P y no Q").

Si nos apegamos a la superveniencia global, esto significa que las propiedades B supervienen lógicamente a los hechos A, si los hechos B acerca del mundo real están implicados por los hechos A. De modo similar, las propiedades B supervienen lógicamente a las propiedades A si no existe ningún mundo concebible con las mismas propiedades A que nuestro mundo pero con propiedades B diferentes. También podemos decir que la superveniencia lógica es válida si, dada la totalidad de hechos-A *A** y cualquier hecho-B *B* acerca de nuestro mundo *W*, "*A**(*W*) *B*(*W*" es verdad en virtud del significado de los términos A y los términos B (donde el significado se entiende como intensión).

Finalmente, si las propiedades B son lógicamente supervenientes a las propiedades A de acuerdo con las intensiones primarias, entonces la implicación de los hechos A a los hechos B será *a priori*. Así, en principio, alguien que conozca todos los hechos A acerca de una situación real podrá averiguar los hechos B acerca de la situación a partir de los hechos A solamente, si posee los conceptos B en cuestión. Este tipo de inferencia puede ser difícil o imposible en la práctica, debido a la complejidad de las situaciones involucradas, pero es al menos posible en principio. Para la superveniencia lógica de acuerdo con las intensiones *secundarias*, los hechos B acerca de una situación pueden también en principio averiguarse a partir de los hechos A, pero sólo *a posteriori*. Los hechos A deberán complementarse con hechos

contingentes acerca del mundo real, ya que estos últimos tendrán un papel en la determinación de las intensiones B involucradas.

Hay, por lo tanto, al menos tres caminos para fundamentar las aseveraciones de superveniencia lógica: estos involucran la concebibilidad, la epistemología y el análisis. Para fundamentar que las propiedades B supervienen lógicamente a las propiedades A podemos 1) argumentar que la instanciación de propiedades A sin instanciación de las propiedades B es inconcebible; 2) argumentar que alguien que está en posesión de los hechos A podría llegar a conocer los hechos B (al menos en casos de superveniencia mediante la intensión primaria), o 3) analizar las intensiones de las propiedades B con suficiente detalle como para que sea evidente que los enunciados B se deducen de los enunciados A en virtud de dichas intensiones solamente. Lo mismo es válido para establecer la no superveniencia lógica. Usaré los tres métodos para defender aseveraciones centrales que involucran a la superveniencia lógica.

No todos estarán convencidos de que las diversas formulaciones de la superveniencia lógica son equivalentes, de modo que cuando argumente en favor de conclusiones importantes acerca de la superveniencia lógica utilizaré versiones de los argumentos que utilicen cada una de las diferentes formulaciones. De este modo se verá que los argumentos son robustos, y que nada depende de una sutil confusión entre las diferentes nociones de superveniencia.

## 5. Casi todo es lógicamente superveniente a lo físico*

En el próximo capítulo argumentaré que la experiencia consciente no superviene lógicamente a lo físico y que, por ende, no puede ser explicarse reductivamente. Una respuesta frecuente es que la experiencia consciente no es la única, y que todo tipo de propiedades no supervienen lógicamente a lo físico. Se sugiere que propiedades tan diversas como la de ser una mesa, la vida y la prosperidad económica no tienen ninguna relación lógica con hechos como átomos, campos electromagnéticos, etc. ¿Es seguro que esos hechos de alto nivel no pueden estar lógicamente implicados por los hechos microfísicos?

En un análisis cuidadoso, pienso que no es difícil ver que esto es erróneo y que los hechos de alto nivel en cuestión son (globalmente) lógicamente supervenientes a lo físico en tanto hechos.[35] La experiencia consciente es casi única por no supervenir lógicamente. La relación entre la conciencia y los hechos físicos es de un tipo diferente de la relación estándar entre los hechos de alto nivel y de bajo nivel.

Hay varios modos de mostrar que la mayoría de las propiedades supervienen lógicamente a las propiedades físicas. Aquí sólo me ocuparé de las propiedades que caracterizan a los *fenómenos naturales*, esto es, aspectos contingentes del mundo que necesitan explicación. La propiedad de ser un ángel podría no supervenir lógicamente a lo físico, pero no tenemos razones para creer en la existencia de los ángeles, de modo que su no superveniencia no tiene por qué preocuparnos. Tampoco me ocuparé de los hechos acerca de entidades abstractas como las entidades matemáticas y las proposiciones; un tema que debe ser tratado separadamente.[36]

Es preciso advertir que cuando afirmo que la mayor parte de las propiedades de alto nivel supervienen a lo físico, no estoy sugiriendo que los hechos y leyes de alto nivel estén implicados por *leyes* microfísicas, o incluso por leyes microfísicas en conjunción con condiciones de borde microfísicas. Esa sería una aseveración fuerte, y aunque podría tener alguna plausibilidad si se la restringe del modo apropiado, todavía no tenemos evidencia para ella. Lo que sostengo es mucho más débil: que los hechos de alto nivel están implicados por los *hechos* microfísicos (quizá junto con leyes microfísicas). Este conjunto enormemente amplio incluye los hechos acerca de la distribución de cada partícula y campo en cada rincón del espaciotiempo: desde los átomos en el gorro de Napoleón a los campos electromagnéticos en el anillo más externo de Saturno. Como veremos, fijar este conjunto de hechos deja poco lugar para que cualquier otra cosa varíe.

Antes de pasar a los argumentos, debo hacer notar algunas razones inofensivas de por qué la superveniencia lógica a lo físico a veces no ocurre. Primero, algunas propiedades de alto nivel no supervienen lógicamente debido a su dependencia de la experiencia consciente. Quizá la experiencia consciente sea parcialmente constitutiva de una propiedad como el amor, por ejemplo. Como veremos, las intensiones primarias (aunque no las secundarias) asociadas con algunas propiedades externas como el color y el calor pueden también depender de cualidades fenoménicas. Si esto es así, entonces el amor y quizás el calor no supervienen lógicamente a lo físico. No debería considerarse que estos constituyen contraejemplos para mi tesis, ya que no introducen ninguna no superveniencia lógica nueva. Quizás el mejor modo de formular la aseveración es decir que todos los hechos supervienen lógicamente a la combinación de hechos físicos y hechos fenoménicos, o que todos los hechos supervienen lógicamente a los hechos físicos *módulo la experiencia consciente*. De modo similar, el hecho de que algunos fenómenos de alto nivel dependan de la experiencia consciente puede dificultar la explicabilidad reductiva,

pero todavía podemos decir que son reductivamente explicables módulo la experiencia consciente.

Segundo, como vimos antes, en la aplicación de algunas intensiones primarias, aunque no en las intensiones secundarias, entra un elemento *indicador*. La intensión primaria de "agua", por ejemplo, es algo así como "el líquido claro y bebible en nuestro ambiente", de modo que si hay $H_2O$ acuoso y XYZ acuoso en el universo real, cuál de ellos calificará como "agua" depende de cuál se encuentre en el ambiente del agente que utiliza el término. En principio, entonces, necesitamos agregar un *centro* que representa la ubicación de un agente respecto de la base de superveniencia en algunos casos. Esto produce superveniencia lógica y explicación reductiva módulo la experiencia consciente y la indicatividad.

Finalmente, los casos en los que los hechos de alto nivel están indeterminados no contradicen la superveniencia lógica. La aseveración sólo dice que si los hechos de alto nivel están determinados, lo están por hechos físicos. Si el propio mundo no basta para fijar los hechos de alto nivel, no podemos esperar que los hechos físicos lo hagan. Algunos podrían sugerir que no hay superveniencia lógica si existen dos teorías de alto nivel igualmente buenas del mundo que difieren en su descripción de los hechos de alto nivel. Una teoría podría sostener que un virus está vivo, por ejemplo, mientras que la otra podría sostener que no lo está; entonces los hechos relativos a la vida no están determinados por los hechos físicos. Sin embargo, este no es un contraejemplo sino un caso en el cual los hechos acerca de la vida están indeterminados. Debido a la indeterminación, estamos libres para legislar los términos de una forma u otra cuando ello sea conveniente. Si los hechos *están* determinados, por ejemplo, si es verdad que los virus están vivos— entonces una de las descripciones es simplemente errónea. De cualquier modo, si los hechos acerca de la situación están determinados, entonces están implicados por los hechos físicos.

Argumentaré en favor de la ubicuidad de la superveniencia lógica utilizando argumentos que recurren a la conceptibilidad, a consideraciones epistemológicas y al análisis de los conceptos involucrados.

*Conceptibilidad.* La superveniencia lógica de la mayoría de los hechos de alto nivel puede verse más fácilmente si se utiliza la conceptibilidad como un test de la posibilidad lógica. ¿Qué clase de mundo podría ser idéntico al nuestro en cada hecho microfísico pero ser biológicamente distinto? Digamos que un uombat tuvo dos hijos

en nuestro mundo. Los hechos físicos acerca de nuestro mundo incluirán hechos acerca de la distribución de cada partícula en el fragmento espaciotemporal correspondiente al uombat y sus hijos, sus ambientes y sus historias evolutivas. Si un mundo compartiese esos hechos físicos con el nuestro, pero no fuera un mundo en el cual el uombat tuvo dos hijos, ¿en qué podría consistir esa diferencia? Un mundo de esta clase parece inconcebible. Una vez que un mundo posible ha sido fijado de modo de que todos esos hechos físicos son iguales, entonces los hechos acerca de ser un uombat y su paternidad quedan automáticamente fijados. Estos hechos biológicos no son la clase de cosa que pueda liberarse de sus apoyos físicos, ni siquiera como posibilidad conceptual.

Lo mismo ocurre para los hechos arquitectónicos, los hechos astronómicos, los hechos conductuales, los hechos químicos, los hechos económicos, los hechos meteorológicos, los hechos sociológicos, etc. Un mundo físicamente idéntico al nuestro, pero en el cual estos tipos de hechos difieran, es inconcebible. Al concebir un mundo idéntico en el nivel microfísico, concebimos un mundo en el cual la localización de cada partícula en el espacio y tiempo es la misma. Se deduce que el mundo tendrá la misma estructura y dinámica macroscópicas que el nuestro. Una vez que todo esto ha sido fijado simplemente no hay espacio para que los hechos en cuestión varíen (aparte, quizá, de cualquier variación debida a diferencias en la experiencia consciente).

Más aún, esta imposibilidad de concepción no parece deberse a ningún límite contingente en nuestra capacidad cognitiva. Un mundo de este tipo es inconcebible *en principio*. Ni siquiera un superser, o Dios, podría imaginarse un mundo de esta clase. Simplemente no hay nada que se puedan imaginar. Una vez que imaginan un mundo con todos los hechos físicos, automáticamente se imaginaron un mundo en el cual rigen todos los hechos de alto nivel. Un mundo físicamente idéntico en el cual los hechos de alto nivel sean falsos es, por lo tanto, lógicamente imposible, y las propiedades de alto nivel en cuestión son lógicamente supervenientes a lo físico.

*Epistemología*. Yendo más allá de las intuiciones de conceptibilidad, podemos notar que si *existiese* un mundo físicamente posible idéntico al nuestro pero biológicamente distinto, esto plantearía problemas epistemológicos radicales. ¿Cómo podríamos saber que no estamos en ese mundo en lugar de en este? ¿Cómo sabríamos que los hechos biológicos en nuestro mundo son como son? Para ver esto, nótese que si estuviésemos en el mundo alternativo, tendría la misma *apariencia* que este. Instanciaría la misma distribución de partículas

que se encuentra en las plantas y animales en este mundo; patrones indistinguibles de fotones se reflejarían de esas entidades; no se revelaría ninguna diferencia, ni siquiera bajo el examen más cercano. Se deduce, entonces, que toda la evidencia externa que poseemos no permite distinguir las posibilidades. Si los hechos biológicos acerca de nuestro mundo no son lógicamente supervenientes, no hay ningún modo de que podamos conocer esos hechos sobre la base de la evidencia externa.

Sin embargo, no existe ningún profundo problema epistemológico acerca de la biología. Todo el tiempo aprendemos hechos biológicos relativos a nuestro mundo basándonos en la evidencia externa; no surge ningún problema escéptico especial. Se deduce, entonces, que los hechos biológicos son lógicamente supervenientes a lo físico. Lo mismo ocurre con los hechos de arquitectura, economía y meteorología. No existe ningún problema escéptico especial acerca de conocer estos hechos sobre la base de la evidencia externa, de modo que deben ser lógicamente supervenientes a lo físico.

Podemos respaldar este punto de vista haciendo notar que en áreas en las que existen problemas epistemológicos, hay un fracaso correlativo de la superveniencia lógica y que, recíprocamente, en áreas en las que la superveniencia lógica falla, existen problemas epistemológicos correlativos.

Más obviamente, existe un problema epistemológico sobre la conciencia, el problema de las otras mentes. Este problema surge porque parece ser lógicamente compatible con toda la evidencia externa que los seres en torno nuestro son seres conscientes, pero también es lógicamente compatible que no lo sean. No tenemos ningún modo de asomarnos al cerebro de un perro, por ejemplo, y observar la presencia o ausencia de la experiencia consciente. El estatus de este problema es controversial, pero la mera existencia *prima facie* del problema es suficiente para rebatir un argumento epistemológico paralelo a los de más arriba en favor de la superveniencia lógica de la conciencia. En cambio, ni siquiera existe un problema *prima facie* con otras biologías u otras economías. Esos hechos son públicamente accesibles, precisamente porque están fijados por los hechos físicos.

(Pregunta: ¿Por qué un argumento similar no nos fuerza a la conclusión de que si la experiencia consciente no superviene lógicamente, entonces no podemos saber ni siquiera acerca de nuestra *propia* conciencia? Respuesta: Porque la experiencia consciente está en el mismo centro de nuestro universo epistémico. Los problemas escépticos acerca de hechos biológicos no supervenientes surgen porque sólo tenemos acceso a los hechos biológicos por medio de la

evidencia externa mediada físicamente; los hechos externos no supervenientes estarían fuera de nuestro alcance epistémico directo. No existe un problema semejante con nuestra propia conciencia.)

Otro famoso problema epistemológico concierne a los hechos acerca de la causalidad. Como argumentó Hume, la evidencia externa sólo nos da acceso a las regularidades en la sucesión de los acontecimientos; no nos da acceso a ningún otro hecho sobre la causalidad. De modo que si la causalidad se interpreta como algo más que la presencia de una regularidad (como supondré que debe ser), no es claro que podamos saber de su existencia. Una vez más, este problema escéptico va de la mano de la no existencia de la superveniencia lógica. En este caso, los hechos acerca de la causalidad no supervienen lógicamente a cuestiones de hechos físicos particulares. Una vez fijados todos los hechos acerca de la distribución de las entidades físicas en el espaciotiempo, es lógicamente posible que todas las regularidades allí incluidas hayan surgido como una gigantesca coincidencia cósmica sin ninguna verdadera causalidad. En una escala más pequeña, dados los hechos particulares acerca de cualquier instancia aparente de causalidad, es lógicamente posible que sea una mera sucesión. Inferimos la existencia de la causalidad mediante una especie de inferencia de la mejor explicación —creer otra cosa sería creer en vastas e inexplicables coincidencias—, pero la creencia en la causalidad no se nos impone del modo directo en que la creencia en la biología lo hace.

Evité los problemas sobre la superveniencia de la causalidad al estipular que, para nuestros propósitos, la base de superveniencia incluye no sólo los hechos físicos particulares sino también todas las leyes físicas. Es razonable suponer que la adición de leyes determine los hechos sobre la causalidad. Pero, por supuesto, existe un problema escéptico acerca de las leyes que es paralelo al problema de la causalidad: considere el problema de Hume sobre la inducción, y la posibilidad lógica de que cualquier aparente ley podría ser una regularidad accidental.

Hasta donde puedo decir, estas dos dificultades agotan los problemas epistemológicos que surgen de la no superveniencia lógica a lo físico. Existen algunos otros problemas epistemológicos que, en cierto sentido, preceden a estos, porque conciernen a la existencia de los propios hechos físicos. Primero tenemos el problema de Descartes acerca de la existencia del mundo externo. Es compatible con nuestra evidencia experiencial que el mundo que pensamos que estamos viendo no exista; quizás estamos alucinando o sólo somos cerebros alojados en tanques. Puede verse que este problema surge precisamente porque los hechos acerca del mundo externo no supervienen

lógicamente a los hechos acerca de nuestra experiencia. (Idealistas, positivistas y otros argumentaron controversialmente que sí lo hacen. Nótese que si se aceptan estos enfoques el problema escéptico se debilita.) También existe un problema epistemológico acerca de las entidades teóricas postuladas por la ciencia: electrones, quarks, etc. Su ausencia sería lógicamente compatible con los hechos directamente observables acerca de los objetos en nuestro ambiente, y algunos, por lo tanto, plantearon dudas escépticas acerca de su existencia. Es posible analizar este problema sobre la base de la no superveniencia lógica de los hechos teóricos a los hechos observacionales. En ambos casos, quizás el mejor modo de mitigar las dudas escépticas sea una forma de inferencia de la mejor explicación, exactamente como en el caso de la causalidad, pero la posibilidad en principio de que estemos equivocados persiste.

De cualquier forma, evito este tipo de problema escéptico presuponiendo el mundo físico y fijando todos los hechos físicos acerca del mundo en la base de superveniencia (en consecuencia, asumo que el mundo externo existe, que existen los electrones, etc.). Si suponemos que estos hechos son conocidos, no hay lugar para dudas escépticas acerca de la mayoría de los hechos de alto nivel, precisamente porque son lógicamente supervenientes. Para decirlo al revés: todas nuestras fuentes de evidencia externa supervienen lógicamente a los hechos microfísicos, de modo que si algún fenómeno no superviene a estos hechos, la evidencia externa no puede suministrarnos ninguna razón para creer en él. Podríamos preguntarnos si es posible postular otros fenómenos mediante la inferencia de la mejor explicación, como hicimos más arriba, para explicar los hechos microfísicos. Este proceso nos lleva desde hechos particulares a leyes subyacentes simples (y por ende produce causalidad), pero luego el proceso parece detenerse. Es parte de la naturaleza de las leyes fundamentales que ellas sean el final de la cadena explicativa (excepto, quizá, para la especulación teológica). Esto nos deja los fenómenos para los que tenemos evidencia *interna* —a saber la experiencia consciente— y nada más. Módulo la experiencia consciente, todos los fenómenos son lógicamente supervenientes a lo físico.

Podemos también hacer una defensa epistemológica de la superveniencia lógica de un modo más directo, argumentando que alguien en posesión de todos los hechos físicos podría en principio llegar a conocer todos los hechos de alto nivel, siempre que posea los conceptos de alto nivel involucrados. Es verdad que nunca podríamos *en la práctica* verificar los hechos de alto nivel a partir del conjunto de hechos microfísicos. La vastedad de este último conjunto es suficiente para desechar esa posibilidad. (Sugiero aun menos que

podríamos realizar una derivación formal; los sistemas formales son irrelevantes por las razones examinadas anteriormente.) Pero, como cuestión de principio, hay diversos modos de mostrar que alguien (¿un superser?) armado solamente con los hechos microfísicos y los conceptos involucrados podría inferir los hechos de alto nivel.

La forma más simple es notar que en principio podríamos construir una gran simulación mental del mundo y observarla con un ojo mental, por así decirlo. Digamos que un hombre lleva un paraguas. A partir de los hechos microfísicos asociados, podríamos directamente inferir hechos acerca de la distribución y composición química de la masa en la vecindad del hombre, y hacer una caracterización estructural de alto nivel del área. Podríamos determinar bastante fácilmente la existencia de un bípedo carnoso varón. Por ejemplo, a partir de la información estructural podríamos notar que hay un organismo sobre dos piernas largas que son responsables de su locomoción, que la criatura tiene una anatomía masculina, etc. Sería claro que transporta algún artefacto que impide que las gotas de agua, predominantes en la vecindad, lo toquen. Podríamos mitigar las dudas de que este artefacto sea realmente un paraguas advirtiendo, a partir de su estructura física, que puede abrirse y cerrarse; a partir de su historia, que esta mañana se encontraba en un paragüero y que originalmente fue hecho en una fábrica con otros de un tipo similar, etc. Las dudas de que el bípedo carnoso sea realmente un ser humano podrían mitigare a través de la composición de su ADN, su historia evolutiva, su relación con otros seres, etc. Sólo necesitamos suponer que el ser posee el concepto involucrado en un grado suficiente como para poder aplicarlo correctamente a instancias (esto es, el ser posee la intensión). Si esto es así, entonces los hechos microfísicos le darán toda la evidencia que necesita para aplicar los conceptos y determinar que realmente hay una persona allí que porta un paraguas.

Lo mismo ocurre para casi cualquier tipo de fenómeno de alto nivel: mesas, vida, prosperidad económica. Conociendo todos los hechos de bajo nivel, un ser puede, en principio, inferir todos los hechos necesarios para determinar si esta es una instancia o no de la propiedad involucrada. Lo que sucede es que se construye un mundo posible compatible con los hechos microfísicos, y los hechos de alto nivel simplemente se leen a partir de ese mundo utilizando la intensión apropiada (ya que los hechos relevantes son invariantes a través de los mundos posibles físicamente idénticos). Por lo tanto, los hechos de alto nivel son lógicamente supervenientes a lo físico.

*Analizabilidad.* Hasta ahora, argumenté que los hechos microfísicos fijan los hechos de alto nivel sin ser muy explícito acerca de los

conceptos de alto nivel involucrados. En cualquier caso específico, sin embargo, esta relación de implicación se basa en la intensión de un concepto. Si los hechos microfísicos implican un hecho de alto nivel, esto se debe a que los hechos microfísicos son suficientes para fijar las características del mundo en virtud de las cuales se aplica la intensión de alto nivel. Esto es, debemos poder *analizar* qué es lo que se requiere para que una entidad satisfaga la intensión de un concepto de alto nivel, al menos en una medida suficiente como para que se pueda ver por qué esas condiciones de satisfacción podrían cumplirse fijando los hechos físicos. Es útil, por lo tanto, estudiar más de cerca las intensiones de los conceptos de alto nivel, y examinar las características del mundo en virtud del cual se aplican.

Existen algunos obstáculos para elucidar estas intensiones y sintetizarlas en palabras. Como vimos antes, las condiciones de aplicación de un concepto suelen estar indeterminadas en diversos sitios. ¿Un objeto con forma de taza pero hecho de tejido es una taza? ¿Un virus computacional está vivo? ¿Una entidad similar a un libro que se materializa aleatoriamente es un libro? Nuestros conceptos ordinarios no dan respuestas directas a estas preguntas En cierto sentido, es una cuestión de estipulación. Por lo tanto, no habrá condiciones de aplicación determinadas que podamos utilizar en el proceso de implicación. Pero, como vimos antes, esta indeterminación refleja precisamente una indeterminación acerca de los propios hechos. En la medida en que la intensión de "taza" es una cuestión de estipulación, los hechos acerca de las tazas también lo son. Lo que cuenta para nuestros propósitos es que la intensión junto con los hechos microfísicos determinan los hechos de alto nivel en la medida en que estos son realmente fácticos. La vaguedad y la indeterminación pueden dificultar la discusión, pero no afectan nada importante en las cuestiones relevantes.

Un problema relacionado es que cualquier análisis conciso de un concepto invariablemente fallará en hacerle justicia. Como hemos visto, los conceptos no suelen tener definiciones precisas. En una primera aproximación, podemos decir que algo es una mesa si tiene una superficie plana horizontal con patas como soporte; pero esto permite la inclusión de demasiadas cosas (¿el monstruo de Frankenstein en zancos?) y omite otras (¿una mesa sin patas, que sobresale de la pared?). Podemos refinar la definición, agregando nuevas condiciones y cláusulas, pero rápidamente nos topamos con los problemas de la indeterminación y, en cualquier caso, el resultado nunca será perfecto. Pero no hay ninguna necesidad de entrar en todos los detalles necesarios para tratar cada caso especial: más allá de un punto, los detalles son sólo más de lo mismo. Si sabemos en

virtud de qué *tipo* de propiedades se aplica la intensión, tendremos suficiente para el caso.

Como vimos antes, no necesitamos una definición de propiedades B en términos de propiedades A para que los hechos A impliquen los hechos B. Los significados están representados fundamentalmente por intensiones, no por definiciones. Aquí, el papel del análisis es simplemente caracterizar las intensiones con suficiente detalle como para que la existencia de una implicación sea evidente. Para este propósito, bastará un análisis aproximado. Las intenciones por lo general se aplican a individuos en un mundo posible en virtud de algunas de sus propiedades y no de otras; el sentido de un análisis de esta clase es determinar en virtud de qué tipo de propiedades se aplica la intensión, y establecer que dichas propiedades son compatibles con la implicación a partir de propiedades físicas.

Un tercer problema surge de la división entre las condiciones de aplicación *a priori* y *a posteriori* de muchos conceptos. Sin embargo, siempre que mantengamos separadas las intensiones primarias y secundarias, esto no será un gran problema. La intensión secundaria asociada con "agua" es algo así como "$H_2O$", lo que, obviamente, es lógicamente superveniente a lo físico. Pero la intensión primaria, algo así como "el líquido claro y bebible en nuestro ambiente" también es lógicamente superveniente, ya que la claridad, potabilidad, y liquidez del agua están implicadas por los hechos físicos.[37] Podemos considerar las cosas de cualquiera de las dos maneras. Como hemos visto, es la intensión primaria la que participa de la explicación reductiva, de modo que es esta la que más nos preocupa. En general, si una intensión primaria *I* es lógicamente superveniente a lo físico, entonces también lo será una intensión secundaria rigidificada *dthat(I),* ya que por lo general consistirá de una proyección de alguna estructura física intrínseca a través de mundos.

Las consideraciones acerca de una necesidad *a posteriori* llevaron a algunos a suponer que no puede haber ninguna implicación lógica desde hechos de nivel inferior a hechos de alto nivel. Típicamente escuchamos cosas como "El agua es necesariamente $H_2O$, pero esa no es una verdad del significado, de modo que no existe ninguna relación conceptual". Pero esta es una simplificación excesiva. Para empezar, la intensión secundaria "$H_2O$" puede verse como parte del significado de "agua" en algún sentido, y por cierto es lógicamente superveniente. Pero aun más importante, la intensión primaria ("el líquido claro y bebible ...") que fija la referencia también es superveniente, quizá módulo la experiencia y la indicatividad. Es precisamente en virtud de que satisface esta intensión que consideramos que el $H_2O$ es agua en primer lugar. Dada la intensión primaria *I*, los

hechos de alto nivel son derivables de un modo no problemático de los hechos microfísicos (módulo la contribución de la experiencia y la indicatividad). La observación kripkeana de que el concepto tiene una mejor representación como *dthat(I)* no afecta en absoluto a esa derivabilidad. Por sí solo, el fenómeno semántico de la rigidificación no marca una diferencia ontológica.

Con estos obstáculos fuera del camino, podemos considerar las intensiones asociadas a los diversos conceptos de alto nivel. En la mayoría de los casos estos pueden caracterizarse en términos funcionales o estructurales, o como una combinación de los dos. Por ejemplo, las clases de cosas relevantes para que algo sea una mesa incluyen 1) que tenga una superficie plana y esté apoyado en patas, y 2) que las personan la usen para sostener diversos objetos. La primera es una condición estructural: esto es, una condición sobre la estructura física intrínseca del objeto. La segunda es una condición funcional: o sea, concierne al papel causal de una entidad, y caracteriza el modo como interactúa con otras entidades. Las propiedades estructurales están claramente implicadas por hechos microfísicos. También lo están las propiedades funcionales en general, aunque esto es levemente menos directo. Estas propiedades dependen de una base de superveniencia mucho más amplia de hechos microfísicos, de modo que los hechos acerca del ambiente de un objeto con frecuencia serán relevantes; y en la medida que tales propiedades se caracterizan por medio de disposiciones (algo es soluble si se *disolvería si* se lo sumergiese en agua), necesitamos entonces recurrir a los contrafácticos. Pero los valores de verdad de esos contrafácticos están fijados por la inclusión de leyes físicas en el antecedente de nuestros condicionales de superveniencia, de modo que esto no es un problema.

Para tomar otro ejemplo, las condiciones sobre la vida se reducen aproximadamente a alguna combinación, entre otras cosas, de la capacidad para reproducirse, adaptarse y metabolizar (como es usual, no necesitamos legislar acerca de su importancia o acerca de todos los otros factores relevantes). Estas propiedades son todas caracterizables funcionalmente, en términos de la relación de una entidad con otras entidades, su capacidad para convertir recursos externos en energía y su capacidad para reaccionar en forma apropiada a su ambiente. Estas propiedades funcionales son todas derivables, en principio, de los hechos físicos. Como es usual, aunque no haya ninguna definición perfecta de la vida en términos funcionales, este tipo de caracterización nos muestra que es una propiedad funcional cuya instanciación puede entonces ser implicada por hechos físicos.

Surge una complicación debido al hecho de que las propiedades funcionales suelen caracterizarse en términos de un papel causal

relativo a otras entidades de alto nivel. Se deduce que la superveniencia lógica de las propiedades depende de la superveniencia lógica a las otras nociones de alto nivel involucradas, donde estas nociones pueden ellas mismas ser caracterizadas funcionalmente. Esto no es, en última instancia, un problema, siempre que los papeles causales eventualmente se traduzcan en propiedades no funcionales: típicamente en propiedades estructurales o fenoménicas. Podría haber una cierta circularidad en la interdefinibilidad de diversas propiedades funcionales: quizá sea parcialmente constitutivo de una grapadora que entrega grapas, y parcialmente constitutivo de las grapas que son entregadas por grapadoras. Esta circularidad puede solucionarse traduciendo los papeles causales de todas las propiedades simultáneamente,[38] siempre que los análisis tengan una parte no circular que se base en última instancia en propiedades estructurales o fenoménicas. (Se podría creer que recurrir a las propiedades fenoménicas va en contra de la superveniencia lógica a lo físico, pero véase más adelante. En cualquier caso, es compatible con la superveniencia lógica módulo la experiencia consciente.)

Muchas propiedades tienen una caracterización relacional en términos de relaciones con el ambiente de una entidad. Por lo general estas relaciones son causales, de modo que las propiedades en cuestión son funcionales, pero esto no siempre es así: considérese la propiedad de estar en el mismo continente que un pato. De modo similar, algunas propiedades dependen de la historia (aunque estas pueden usualmente interpretarse de modo causal); para ser un canguro, una criatura debe tener predecesores apropiados. De cualquier modo, estas propiedades no plantean ningún problema para la superveniencia lógica, ya que los hechos históricos y ambientales relevantes estarán fijados por los hechos físicos globales.

Aun un hecho social complejo como "En los años cincuenta existió prosperidad económica"[39] puede caracterizarse principalmente en términos funcionales y, por lo tanto, considerarse que está implicado por hechos físicos. Un análisis completo sería muy complicado y se vería dificultado por la vaguedad de la noción de prosperidad, pero, para tener una idea de cómo sería, podemos preguntarnos por qué decimos que *hubo* prosperidad económica en los años cincuenta. En primera aproximación, porque la tasa de empleo era alta, las personas podían comprar cantidades inusualmente grandes de bienes, la tasa de inflación era baja, había un gran desarrollo en viviendas, etc. A su vez podemos hacer un análisis grueso de la noción de vivienda (el tipo de lugar en donde las personas duermen y comen), de empleo (trabajo organizado que procura una recompensa) y de las nociones monetarias (dinero sería analizable aproximadamente en

términos de su capacidad sistemática para ser intercambiado por otros objetos, y su valor será analizable en términos de cuánto obtenemos en el intercambio). Todos estos análisis están ridículamente sobresimplificados, pero el punto es bastante claro. Estas son, en general, propiedades funcionales que pueden estar implicadas por hechos físicos.

Muchos han sido escépticos respecto de la posibilidad de los análisis conceptuales. Frecuentemente, esto se debió a razones que no son relevantes para mis argumentos: debido a la indeterminación en nuestros conceptos, por ejemplo, o porque carecen de definiciones precisas. A veces, el escepticismo puede haber surgido por razones más profundas. Sin embargo, si lo que dije anteriormente en este capítulo es correcto, y si los hechos físicos acerca de un mundo posible determinan los hechos de alto nivel, deberíamos *esperar* estar en condiciones de analizar la intensión del concepto de alto nivel en cuestión, al menos en una buena aproximación, para ver cómo su aplicación puede estar determinada por hechos físicos. Esto es lo que intenté hacer en los ejemplos dados aquí. Otros ejemplos pueden manejarse de un modo similar.[40]

No propugno un programa para realizar estos análisis en general. Los conceptos son demasiado complejos y rebeldes para que resulte de mucha utilidad, y cualquier análisis explícito suele ser una pálida sombra de la cosa real. Lo que importa es la cuestión básica de que la mayoría de los conceptos de alto nivel no son nociones primitivas inanalizables. En general, son analizables en la medida en que pueda considerarse que sus intensiones especifican propiedades funcionales o estructurales. Es en virtud de esta analizabilidad que los hechos de alto nivel son, en principio, derivables de los hechos microfísicos y explicables reductivamente en términos de hechos físicos.

## Algunos casos problemáticos

Podría pensarse que existen algunos tipos de propiedades que presentan dificultades particulares para la superveniencia lógica y, por lo tanto, para la explicación reductiva. Examinaremos algunos de estos casos, prestando particular atención a la cuestión de si el fenómeno asociado plantea para la explicación reductiva problemas análogos a los planteados por la conciencia. Me parece que, con un par de posibles excepciones, no surgen aquí nuevos problemas significativos.

*Propiedades dependientes de la conciencia.* Como ya se expuso, las intensiones primarias de algunos conceptos involucran una rela-

ción con la experiencia consciente. Un ejemplo obvio es la propiedad de rojez, considerada como propiedad de objetos externos. Según algunas concepciones, la intensión primaria asociada con la rojez requiere que para que algo sea rojo debe ser el tipo de cosa que tiende a causar experiencias de color rojo bajo condiciones apropiadas.[41] De modo que en su intensión primaria, la rojez no es lógicamente superveniente a lo físico, aunque superviene módulo la experiencia consciente. Por otro lado, su intensión secundaria casi seguramente es superveniente. Si resulta que en el mundo real, el tipo de cosa que tiende a causar experiencias de color rojo es la reflectancia de una cierta superficie, entonces los objetos con esa reflectancia son rojos aun en mundos en los que no haya seres conscientes para verlos. La rojez se identifica *a posteriori* con esa reflectancia, que es sólo lógicamente superveniente a lo físico.

Vimos con anterioridad que la no superveniencia lógica de una intensión primaria está asociada a un fracaso de la explicación reductiva. Entonces, ¿falla la explicación reductiva para la rojez? La respuesta es sí en un sentido débil. Si se interpreta la rojez como la tendencia a causar experiencias de color rojo, entonces si la experiencia no es reductivamente explicable, tampoco lo es la rojez. Pero podemos acercarnos. Podemos notar que una cierta cualidad física causa experiencias de color rojo; e incluso podemos explicar la relación causal entre la cualidad y los *juicios* de rojo. Es sólo el paso final de la experiencia lo que no tiene explicación. En la práctica, nuestras restricciones sobre la explicación son lo suficientemente débiles como para que este tipo de cosas cuente. Para explicar un fenómeno cuya referencia está fijada por alguna experiencia, no requerimos una explicación de la experiencia. De otro modo deberíamos esperar durante mucho tiempo.

Lo mismo ocurre para fenómenos como el calor, la luz y el sonido. Aunque sus intensiones secundarias determinan propiedades estructurales (movimiento molecular, la presencia de fotones, ondas en el aire), sus intensiones primarias involucran una relación con la experiencia consciente: el calor es lo que causa sensaciones de calor, la luz causa experiencias visuales, etc. Pero, como Nagel (1974) y Searle (1992) hicieron notar, no necesitamos una explicación de las sensaciones de calor cuando explicamos el calor. La explicación módulo la experiencia es suficiente.

Otras propiedades dependen más directamente de la experiencia consciente, en el sentido de que la experiencia no sólo tiene un papel en la determinación de la referencia sino que también es parte constitutiva de la noción *a posteriori*. La propiedad de estar parado al lado de una persona consciente es un ejemplo obvio. Según algunas

concepciones, propiedades mentales como el amor y la creencia, aunque no son en sí mismas propiedades fenoménicas, dependen conceptualmente de la existencia de la experiencia consciente. Si esto es así, entonces en un mundo sin conciencia, estas propiedades no estarían ejemplificadas. Estas propiedades, por lo tanto, no son lógicamente supervenientes ni siquiera *a posteriori*, y la explicación reductiva falla de un modo aun más contundente que en los casos de arriba. Pero son lógicamente supervenientes y reductivamente explicables módulo la experiencia consciente, de modo que no surge aquí ningún nuevo fracaso de la explicación reductiva.

*Intencionalidad.* Vale la pena considerar separadamente el estado de intencionalidad, ya que a veces se considera que plantea problemas análogos a los que surgen debido a la conciencia. Es plausible, sin embargo, que cualquier fracaso de las propiedades intencionales en supervenir lógicamente se derive de la no superveniencia de la conciencia. Como hice notar en el capítulo 1, no parece haber ningún mundo concebible que sea física y fenoménicamente idéntico al nuestro, pero en el cual los contenidos intencionales difieran.[42] Si la fenomenología es parcialmente constitutiva del contenido intencional, como algunos filósofos sugieren, entonces las propiedades intencionales pueden no supervenir lógicamente a lo físico, pero sí lo harán módulo la experiencia consciente. La aserción de que la conciencia es parcialmente constitutiva del contenido es controversial, pero de cualquier forma hay pocas razones para creer que la intencionalidad falle en supervenir de un modo separado y no derivado.

Dejando de lado los aspectos fenomenológicos, es mejor considerar las propiedades intencionales como una clase de construcción de tercera persona en la explicación de la conducta humana, por lo que debería, entonces, ser analizable en términos de conexiones causales con la conducta y el ambiente. Si esto es así, las propiedades intencionales son entonces lógicamente supervenientes a lo físico. Lewis (1974) hace un intento detallado de explicitar la implicación de los hechos físicos con los hechos intencionales mediante un apropiado análisis funcional. Puede interpretarse que concepciones más recientes de la intencionalidad, como las de Dennett (1987), Dretske (1981), y Fodor (1987), contribuyen al mismo proyecto. Ninguno de estos análisis es totalmente convincente, pero podría ocurrir que un descendiente más sofisticado cumpla con el trabajo. No existe ningún argumento que sea análogo a los argumentos en contra de la superveniencia de la conciencia que muestre que la intencionalidad no puede supervenir lógicamente a las propiedades físicas y fenoméni-

cas.[43] Los argumentos de conceptibilidad indican que las propiedades intencionales deben ser lógicamente supervenientes a estas en la medida en que dichas propiedades se instancien, y los argumentos epistemológicos nos llevan a una conclusión similar. De este modo, no existe ningún problema *ontológico* independiente de la intencionalidad.

*Propiedades morales y estéticas.* Suele sostenerse que no existe ninguna conexión *conceptual* entre las propiedades físicas y las propiedades morales y estéticas. De acuerdo con Moore (1922), nada en el *significado* de nociones como "bondad" permite que los hechos relativos a la bondad deban estar implicados con hechos físicos. De hecho, Moore sostenía que no existe ninguna conexión conceptual entre los hechos *naturales* y los hechos morales, donde lo natural puede incluir lo mental además de lo físico (de modo que la superveniencia módulo la experiencia consciente no nos ayuda aquí). ¿Significa esto que las propiedades morales son tan problemáticas como la experiencia consciente?

Existen dos diferencias, sin embargo. Primero, no parece haber un mundo concebible que sea naturalmente idéntico al nuestro pero moralmente distinto, de modo que es improbable que los hechos morales sean nuevos hechos en algún sentido fuerte. Segundo, los hechos morales no son fenómenos que se nos impongan. Si se nos presiona, podemos directamente negar que los hechos morales existan. Esto refleja la estrategia adoptada por los antirrealistas morales como Blackburn (1971) y Hare (1984). Estos antirrealistas argumentan que debido a que los hechos morales no están implicados por hechos naturales y no parecen ser nuevos hechos "extraños", no tienen ninguna existencia objetiva y la moralidad debe ser relativizada en una construcción o proyección de nuestro aparato cognitivo. No es posible adoptar la misma estrategia para las propiedades fenoménicas, cuya existencia se nos impone.

Para las propiedades morales, existen al menos dos alternativas razonables disponibles. La primera es un antirrealismo de alguna clase, que quizá relativice los "hechos morales objetivos" en "hechos morales subjetivos",[44] o adopte un enfoque en el que el discurso moral no afirma hechos en absoluto. La segunda es afirmar que existe una conexión *a priori* entre los hechos naturales y los hechos morales, una que (contra Moore) puede considerarse que es válida en virtud de un análisis y explicación de conceptos morales. Si un concepto como "bueno" determina una extensión primaria no indicadora, entonces se deduce la segunda posición: tendremos una función *a priori* desde los mundos naturalmente especificados hasta los hechos mora-

les. Si sólo determina una intensión primaria indicadora, o si diferentes sujetos pueden asociar distintas intensiones primarias al concepto, o si no determina ninguna intensión primaria en absoluto, entonces se deducirá una versión de la primera posición.

A veces se adoptan algunas otras posiciones, pero ninguna parece defendible. Moore sostenía que existe una conexión *a priori* no conceptual entre los hechos naturales y los morales que obtenemos a través de una misteriosa facultad de "intuición moral", pero este enfoque es por lo general rechazado (resulta difícil determinar qué es lo que podría fundamentar la verdad o falsedad de una intuición de este tipo). Una posición según la cual las propiedades morales supervienen debido a un vínculo nomológico fundamental parece fuera de cuestión, ya que no existe ningún mundo concebible en el cual los hechos naturales sean los mismos que en el nuestro pero en el cual los hechos morales sean diferentes. Una posición popular entre los realistas morales contemporáneos (véase, por ejemplo, Boyd, 1988; Brink, 1989) es que los hechos morales supervienen a los hechos naturales con necesidad *a posteriori*; esto es, supervienen de acuerdo con las intensiones secundarias pero no las primarias de los conceptos morales. Esta posición es difícil de sostener, sin embargo, dado que incluso las equivalencias *a posteriori* deben basarse en una determinación *a priori* de la referencia. Aun cuando el agua es $H_2O$ *a posteriori*, los hechos acerca del agua se deducen *a priori* de los hechos microfísicos. De modo similar, si los conceptos morales tienen una intensión primaria y si los mundos centrados naturalmente idénticos son moralmente idénticos, parecería deducirse un vínculo *a priori* de los hechos naturales con los hechos morales (Horgan y Timmons [1992a; 1992b] realizan una crítica de este tipo).

Las propiedades estéticas pueden tratarse de un modo similar. Un tratamiento antirrealista es quizás aun más plausible aquí. En un análisis final, aunque existen cuestiones conceptuales interesantes acerca de cómo deberían considerarse los dominios moral y estético, estos no plantean problemas metafísicos y explicativos comparables a los ocasionados por la experiencia consciente.

*Nombres.* Según muchas concepciones (por ejemplo, Kaplan, 1989), no existe ningún análisis asociado con un nombre como "Rolf Harris", que simplemente seleccione su referente en forma directa. ¿Significa esto que la propiedad de ser Rolf Harris no superviene lógicamente a lo físico? No hay ningún problema con la superveniencia de la intensión secundaria (por ejemplo, Rolf podría ser la persona concebida a partir de un esperma y huevo determinados en todo mundo posible), pero podría pensarse que la falta de una intensión

primaria plantea problemas a la explicación reductiva. Sin embargo, es plausible que aunque no exista ninguna intensión primaria compartida a través de la comunidad, toda utilización individual del nombre tiene una intensión primaria asociada. Cuando utilizo el nombre "Rolf Harris" existe *algún* modo sistemático en el cual su referente depende del modo como el mundo resulta; para mí, la intensión primaria podría ser algo así como "el hombre llamado 'Rolf Harris' que golpea latas de pintura y que tiene conmigo la relación causal apropiada".[45] Una intensión de esta clase será lógicamente superveniente. En lugar de justificar esto en detalle, sin embargo, es más fácil notar que cualquier ausencia de superveniencia lógica no estará acompañada por un misterio explicativo. La propiedad de ser Rolf Harris no constituye un fenómeno que necesite una explicación, sí en cambio una explicitación. Lo que debe explicarse es la existencia de una persona llamada "Rolf Harris" que golpea latas de pintura, etc. Estas propiedades ciertamente supervienen, y son explicables en principio del modo usual.

*Expresiones indicadoras*. La determinación de la referencia de muchos conceptos, desde "agua" a "mi perro", involucra un elemento indicador. La referencia de estas nociones se fija sobre la base de los hechos físicos y de un "hecho indicador" relativo al agente que representa la ubicación del agente que utiliza el término en cuestión. Un hecho de esta clase está determinado para cualquier agente dado, de modo que la fijación de la referencia está determinada. La superveniencia y la explicación tienen éxito módulo ese hecho indicador.

¿La indicatividad plantea algún problema para la explicación reductiva? Para hablantes arbitrarios, quizá no, ya que el "hecho" en cuestión puede ser relativizado. Pero para mí mismo, no es tan fácil. El hecho indicador expresa algo muy sobresaliente acerca del mundo tal como lo encuentro: que David Chalmers soy *yo*. ¿Cómo se puede explicar este hecho aparentemente primitivo? Por cierto, ¿hay aquí realmente un hecho que debe ser explicado, y no una tautología? La cuestión es extraordinariamente difícil de captar, pero me parece que aunque el término indicador no es un hecho objetivo acerca del mundo, es un hecho acerca del mundo tal como yo lo encuentro, y es el mundo tal como yo lo encuentro lo que necesita una explicación. La naturaleza de un indicador primitivo es bastante oscura, sin embargo, y es muy poco claro cómo se lo podría explicar.[46] (Por supuesto, podemos dar una explicación reductiva de por qué la emisión de David Chalmers "Yo soy David Chalmers" es verdadera. Pero este hecho no indicador parece ser bastante diferente del hecho indicador de que yo soy David Chalmers.)

Es tentador considerar a la conciencia. Pero mientras que una explicación de la conciencia podría producir una explicación de "puntos de vista" en general, es difícil ver cómo podría explicar por qué un punto de vista aparentemente arbitrario de estos es el *mío*, a menos de que el solipsismo sea verdad. Podríamos vernos obligados a tomar el hecho indicador como primitivo. Si esto es así, entonces tenemos un fracaso de la explicación reductiva distinto de y análogo al fracaso de la conciencia. Sin embargo, el primer fracaso es menos preocupante que el de la conciencia, debido a que el hecho inexplicado es muy "tenue" en comparación con los hechos acerca de la conciencia en toda su gloria. Admitir este hecho indicativo primitivo requeriría mucho menos revisión de nuestra cosmovisión materialista que admitir hechos irreducibles acerca de la experiencia consciente.

*Hechos negativos*. Como vimos antes, ciertos hechos que involucran existenciales negativos y cuantificadores universales no están lógicamente determinados por los hechos físicos, o incluso por cualquier conjunto de hechos localizados. Considérense los siguientes hechos acerca del mundo: no hay ángeles; Don Bradman es el mejor jugador de criquet; todos los seres vivos se basan en el ADN. Todos estos hechos podrían ser falsificados, de un modo consistente con todos los hechos físicos acerca de nuestro mundo, si simplemente agregamos alguna nueva sustancia no física: ángeles que juegan al criquet hechos de ectoplasma, por ejemplo. Ni siquiera la adición de hechos acerca de la experiencia consciente o la indicatividad puede ayudarnos aquí.[47]

¿Significa esto que esos hechos no son reductivamente explicables? Así parece, en la medida en que no haya ninguna explicación física de por qué no existe ninguna sustancia no física extra en nuestro mundo. Esto es ciertamente un hecho ulterior. El mejor modo de manejar esta situación es introducir un hecho de segundo orden que dice acerca del conjunto de hechos básicos particulares, sean microfísicos, fenoménicos, indicadores, o los que sean: *Eso es todo*. Este hecho dice que todos los hechos particulares acerca del mundo están incluidos en o implicados por el conjunto dado de hechos. De este hecho de segundo orden, en conjunción con todos los hechos básicos particulares, se deducirán todos los hechos negativos.

Esto no constituye un fracaso muy serio de la explicación reductiva. Es de suponer que existirá un hecho verdadero "Eso es todo" de este tipo en cualquier mundo, y un hecho de esta clase nunca estará implicado con los hechos particulares. Simplemente expresa la naturaleza limitada de nuestro mundo, o de cualquier mundo. Es un modo económico de llevar a nuestra captación todos los hechos existenciales negativos o universalmente cuantificados.

*Leyes físicas y causalidad.* En las concepciones más plausibles de las leyes físicas, estas no son lógicamente supervenientes a los hechos físicos, considerados como una colección de hechos particulares acerca de la historia espaciotemporal del mundo. Se puede ver esto notando la posibilidad lógica de un mundo físicamente indiscernible del nuestro a través de toda su historia espaciotemporal, pero con leyes diferentes. Por ejemplo, podría ser una ley de ese mundo que cuando se reúnen en el vacío doscientas toneladas de oro puro, este se transmutará en plomo. En cualquier otro aspecto sus leyes son idénticas, con modificaciones menores cuando sea necesario. Ocurre que en la historia espaciotemporal de nuestro mundo, nunca se reunieron en el vacío doscientas toneladas de oro. Se deduce que nuestro mundo y el otro mundo tienen historias idénticas, pero sin embargo sus leyes difieren.

Argumentos como este sugieren que las leyes de la naturaleza no son lógicamente supervenientes a la colección de hechos físicos particulares.[48] Mediante argumentos similares podemos ver que una conexión causal entre dos sucesos es algo que va más allá de una regularidad entre ellos. Las personas que sostienen diversos enfoques humeanos ponen en duda estas conclusiones, pero me parece que tienen aquí el peor de los argumentos.[49] Hay algo irreducible en la existencia de las leyes y la causalidad.

Evité estos problemas en otros lados al incluir las leyes físicas en la base de superveniencia, pero esto pasa por encima del problema metafísico en lugar de resolverlo. Es verdad que las leyes y la causalidad llevan a un fracaso menos significativo de la explicación reductiva que la conciencia. Las propias leyes y relaciones causales se postulan para explicar fenómenos físicos existentes, principalmente las múltiples regularidades presentes en la naturaleza, mientras que la conciencia es un explanandum primitivo. Sin embargo, la existencia de hechos irreducibles ulteriores de este tipo plantea profundas preguntas acerca de su naturaleza metafísica. Aparte de la experiencia consciente y, quizá, de la indicatividad, estos son los únicos otros hechos ulteriores en los que tenemos razones para creer. Es natural especular que estas dos clases no supervenientes, la conciencia y la causalidad, puedan tener una estrecha relación metafísica.

## Recapitulación

La posición en la que quedamos es que casi todos los hechos supervienen lógicamente a los hechos físicos (incluyendo las leyes físicas), con las posibles excepciones de la experiencia consciente, la indicatividad y los hechos existenciales negativos. Para presentar las cosas de otro modo, podemos decir que los hechos acerca del mundo se

agotan en 1) hechos físicos particulares, 2) hechos acerca de la experiencia consciente, 3) leyes de la naturaleza, 4) un hecho de segundo orden "Eso es todo" y, quizá, 5) un hecho indicativo acerca de mi ubicación. (Los últimos dos son menores en comparación con los otros, y la condición del último de ellos es dudosa, pero lo incluyo aquí por razones de completitud.) Módulo la experiencia consciente y la indicatividad, parece que todos los hechos positivos son lógicamente supervenientes a lo físico. Establecer esto en forma concluyente requeriría un examen más detallado de todo tipo de fenómenos, pero lo que hemos visto sugiere que la conclusión es razonable. Podemos sintetizar las situaciones ontológicas y epistemológicas mediante un par de fábulas. Quizás haya una pizca de verdad en la forma de estas historias, si no en los detalles.

*Mito de la creación.* Para crear el mundo, todo lo que Dios tuvo que hacer fue fijar los hechos recién mencionados. Para una máxima economía de esfuerzo, primero fijó las leyes de la naturaleza: las leyes de la física y todas las leyes que relacionan lo físico con la experiencia consciente. Luego, fijó las condiciones limitativas: quizás un corte temporal de hechos físicos y los valores en un generador de números aleatorios. Estos se combinaron con las leyes para fijar los hechos físicos y fenoménicos restantes. Finalmente, decretó "Eso es todo".

*Mito epistemológico.* Al principio, sólo tengo hechos acerca de mi experiencia consciente. Desde aquí, infiero hechos acerca de objetos de tamaño intermedio en el mundo, y luego hechos microfísicos. De las regularidades en estos hechos, infiero leyes físicas, y por lo tanto nuevos hechos físicos. De las regularidades entre mi experiencia consciente y los hechos físicos, infiero leyes psicofísicas, y por lo tanto hechos acerca de la experiencia consciente en otros. Parece que llevé el proceso abductivo hasta donde puede llegar, de modo que hipotetizo: eso es todo. El mundo es mucho más vasto de lo que en un momento pareció, de modo que distingo las experiencias conscientes originales como *mías.*

Nótese el orden muy diferente involucrado en ambas perspectivas. Casi se podría decir que la epistemología recapitula la ontología al revés. Nótese también que parece estar más allá de los poderes de Dios fijar mi hecho indicativo. Esta podría ser una razón más para mantenerse escéptico sobre ese hecho.

La superveniencia lógica de la mayoría de los fenómenos de alto nivel es una conclusión que no ha sido tan ampliamente aceptada como podría haberlo sido, aun entre aquellos que plantean la superveniencia. Aunque no suele discutirse la cuestión, muchos desconfían de invocar que la modalidad conceptual es relevante para las relacio-

nes de superveniencia. Hasta donde puedo decir se ha formulado un cierto número de razones para esta vacilación, ninguna de las cuales resulta demasiado convincente.

Primero, el problema con los mundos lógicamente posibles y físicamente idénticos con sustancia no física extra (ángeles, ectoplasma) llevó a algunos a suponer que las relaciones de superveniencia no pueden ser lógicas (Haugeland, 1982; Petrie, 1987); pero hemos visto cómo corregir este problema. Segundo, muchos supusieron que las consideraciones acerca de la necesidad *a posteriori* demuestran que los significados no pueden asegurar las relaciones de superveniencia (Brink, 1989; Teller, 1984); pero hemos visto que las relaciones de superveniencia basadas en una necesidad *a posteriori* pueden considerarse una variedad de superveniencia lógica. Tercero, existe un escepticismo generalizado acerca de la noción de verdad conceptual, que proviene de Quine; pero hemos visto que este es un falso camino. Cuarto, las preocupaciones relativas a la "reducibilidad" llevaron a algunos a suponer que la superveniencia no es en general una relación conceptual (Hellman y Thompson, 1975); pero no es claro que existan buenos argumentos en contra de la reducibilidad que sean también buenos argumentos en contra de la superveniencia lógica. Quinto, a veces se invoca el propio fenómeno de la experiencia consciente para demostrar que, en general, las relaciones de superveniencia no pueden ser lógicas (Seager, 1988); pero hemos visto que la experiencia consciente es casi única en su no superveniencia lógica. Finalmente, suele sostenerse sin justificación que las relaciones de superveniencia no son lógicas en general, supuestamente como algo que cualquier persona razonable debe creer (Bacon, 1986; Heil, 1992).[50]

Es plausible que toda relación de superveniencia a lo físico de una propiedad de alto nivel sea en última instancia 1) una relación de superveniencia lógica de la variedad primaria o secundaria, o 2) una relación de superveniencia natural contingente. Si ninguna de estas es válida en el caso de alguna aparente relación de superveniencia, entonces tenemos buenas razones para creer que no existen hechos de alto nivel objetivos del tipo en cuestión (como ocurre, quizá, para los hechos morales). Argumentaré además, en el capítulo 4, que no existe ninguna variedad profunda de superveniencia intermedia entre la lógica y la natural.

Esto proporciona una imagen explicativa unificada, en principio. Casi todos los fenómenos son reductivamente explicables en el sentido débil formulado antes, excepto la experiencia consciente y, quizás, la indicatividad, junto con los hechos y leyes microfísicas de base, que deben ser considerados fundamentales.

Vale la pena que nos tomemos un momento para responder una pregunta formulada por Blackburn (1985) y Horgan (1993): ¿Cómo explicamos las propias relaciones de superveniencia? En el caso de una relación de superveniencia lógica basada en la intensión primaria de un concepto, esto significa simplemente realizar un análisis apropiado del mismo, quizás en términos funcionales o estructurales, y notar que su referencia es invariante a través de mundos físicamente idénticos. Aquí el condicional de superveniencia es en sí mismo una verdad conceptual *a priori*. Para una relación de superveniencia lógica basada en una intensión secundaria, la superveniencia puede explicarse notando que la intensión primaria del concepto selecciona algún referente del mundo actual que se proyecta (por rigidificación) de una manera invariante a través de mundos físicamente idénticos. Todo lo que necesitamos aquí como explicación es un análisis conceptual *a priori* combinado con hechos contingentes acerca del mundo real.[51] Por otro lado, una relación de superveniencia meramente natural será una ley contingente. En el mejor de los casos, será explicable en términos de leyes más fundamentales; en el peor de los casos, la propia ley de superveniencia será fundamental. En cualquier caso, explicamos ciertas regularidades en el mundo invocando leyes fundamentales, tal como lo hacemos en física, y, como siempre, la explicación debe detenerse en las leyes fundamentales. La superveniencia meramente natural es ontológicamente costosa, como hemos visto, de modo que es afortunado que la superveniencia lógica sea la regla y la superveniencia natural la excepción.

# PARTE II

# LA IRREDUCTIBILIDAD DE LA CONCIENCIA

# 3

# ¿Puede la conciencia explicarse reductivamente?

## 1. ¿Es la conciencia lógicamente superveniente a lo físico?

Casi todo en el mundo puede explicarse en términos físicos; es natural esperar que también la conciencia pueda explicarse de ese modo. Sin embargo, en este capítulo argumentaré que la conciencia escapa a la red de la explicación reductiva. Nunca una explicación enunciada totalmente en términos físicos podrá dar cuenta del surgimiento de la experiencia consciente. Esto podría parecer una conclusión negativa, pero lleva a algunas poderosas consecuencias positivas que plantearé en capítulos posteriores.

Para argumentar en contra de la explicación reductiva, debemos mostrar que la conciencia no es lógicamente superveniente a lo físico. En principio, debemos mostrar que no superviene *globalmente*, esto es, que todos los hechos microfísicos en el mundo no implican los hechos acerca de la conciencia. En la práctica, sin embargo, es más fácil aplicar la argumentación a la superveniencia *local*, razonando que en un individuo, los hechos microfísicos no implican los hechos sobre la conciencia. En lo que respecta a la conciencia, la superveniencia local y global se sostienen juntas o caen juntas, de modo que no es muy importante de qué forma desarrollemos el argumento: si la conciencia es superveniente, casi seguro que superviene localmente. Si esto se pone en duda, sin embargo, todos los argumentos pueden formularse en el nivel global realizando modificaciones evidentes.

¿Cómo podemos argumentar que la conciencia no es lógicamente superveniente a lo físico? Hay diversos modos. Podemos pensar acerca de lo que es concebible, y argumentar directamente en favor de

la posibilidad lógica de una situación en la cual los hechos físicos son los mismos pero los hechos acerca de la experiencia son diferentes. Podemos recurrir a la epistemología, y argumentar que el tipo correcto de vínculo entre el conocimiento de los hechos físicos y el conocimiento de la conciencia está ausente. Podemos también recurrir directamente al concepto de conciencia, y argumentar que no existe ningún análisis del concepto que pueda justificar una implicación de lo físico a lo fenoménico. En lo que sigue formularé argumentos que utilizan las tres estrategias. Los dos primeros son esencialmente argumentos a partir de la conceptibilidad, los dos siguientes son argumentos a partir de la epistemología y el quinto es un argumento a partir del análisis. Existe una cierta redundancia entre los cinco, pero juntos forman una sólida defensa.

También podemos hacer las cosas de un modo más directo, y argumentar en contra de la explicación reductiva sin recurrir en forma explícita a la superveniencia lógica. He seguido ese camino en otro lugar: aquí realizaré el análisis más detallado que hace una consideración más completa de la cuestión. De cualquier manera, el caso en contra de la explicación reductiva y la crítica de las concepciones reductivas existentes (del apartado 2 en adelante) debería tener sentido aun sin este análisis. Algunos lectores podrían preferir en una primera lectura proceder directamente a ese lugar.

(Una nota técnica: El peso en este capítulo se encuentra en la argumentación de que no existe ninguna implicación *a priori* de los hechos físicos a los hechos fenoménicos. El tipo de necesidad que define la relación de superveniencia relevante es la versión *a priori* de necesidad lógica, donde las intensiones primarias son fundamentales. Como vimos en el capítulo 2, esta es la relación relevante en las cuestiones acerca de la explicación; las cuestiones de necesidad *a posteriori* pueden dejarse de lado. En el próximo capítulo, las cuestiones de ontología, en lugar de las vinculadas con la explicación, serán las fundamentales, y argumentaré separadamente que no hay ninguna conexión necesaria *a posteriori* entre los hechos físicos y los hechos fenoménicos.)

### Argumento 1: La posibilidad lógica de los zombis

El modo más obvio (aunque no el único) de investigar la superveniencia lógica de la conciencia es considerar la posibilidad lógica de un *zombi*: alguien o algo físicamente idéntico a mí (o a cualquier otro ser consciente), pero que carece por completo de experiencias conscientes.[1] En el nivel global, podemos considerar la posibilidad lógica de un *mundo zombi*: un mundo físicamente idéntico al nuestro, pero

**Calvin y Hobbes** **por Bill Watterson**




*Figura 3.1.* Calvin y Hobbes sobre los zombis. (Calvin and Hobbes © Watterson. Distribuido por Universal Press Syndicate. Reproducido con autorización. Todos los derechos reservados.)

en el cual no existen las experiencias conscientes. En un mundo de este tipo, todos son zombis.

Consideremos entonces a mi gemelo zombi. Esta criatura es idéntica a mí molécula por molécula e idéntica en todas las propiedades de bajo nivel postuladas por una física terminada, pero carece por completo de experiencia consciente. (Algunos podrían preferir llamar a un zombi "eso", pero utilizaré el pronombre personal; me he encariñado bastante con mi gemelo zombi.) Para fijar ideas, podemos imaginar que en este momento estoy mirando a través de la ventana, y experimento algunas agradables sensaciones de verde de los árboles que veo afuera, tengo experiencias gustativas placenteras por estar masticando una barra de chocolate y siento una sensación de dolor apagado en mi hombro derecho.

¿Qué ocurre con mi gemelo zombi? Es físicamente idéntico a mí y podemos también suponer que está inmerso en un ambiente idéntico. Con seguridad será idéntico a mí *funcionalmente*: procesará la misma información, reaccionará de un modo similar a las entradas, sus configuraciones internas se modificarán en forma apropiada y como resultado su conducta será indistinguible de la mía. Será *psicológicamente* idéntico a mí, en el sentido desarrollado en el capítulo 1. Percibirá los árboles afuera, en el sentido funcional, y degustará el chocolate, en el sentido psicológico. Todo esto se deduce lógicamente del hecho de que es físicamente idéntico a mí, en virtud de los análisis funcionales de las nociones psicológicas. Incluso será "consciente" en los sentidos funcionales descriptos antes: estará despierto, será capaz de informar del contenido de sus estados internos, capaz de concentrar su atención en diversos lugares, etc. Sólo que nada de este funcionamiento estará acompañado por alguna

experiencia consciente real. No habrá ninguna experiencia fenoménica. No existe una experiencia de ser como un zombi.

Este tipo de zombi es bastante diferente de los zombis que se pueden ver en las películas de Hollywood, que suelen poseer significativas discapacidades funcionales (fig. 3.1). El tipo de conciencia que los zombis de Hollywood muy obviamente no tienen es una versión psicológica de la misma: típicamente tienen poca capacidad de introspección y carecen de una capacidad refinada de controlar voluntariamente la conducta. Puede o no haber ausencia de la conciencia fenoménica; como Block (1995) señala, es razonable suponer que hay algo que es una degustación cuando comen a sus víctimas. Los podemos llamar *zombis psicológicos*; yo estoy interesado más bien en los *zombis fenoménicos*, que son física y funcionalmente idénticos a nosotros, pero carecen de la capacidad de experimentar. (Quizá no sea sorprendente que los zombis fenoménicos no hayan sido populares en Hollywood, ya que habría problemas obvios con su descripción.)

La idea de zombis tal como los describí es extraña. Para empezar, es improbable que los zombis sean naturalmente posibles. En el mundo real, es probable que cualquier réplica mía sea consciente. Por esta razón, es muy natural imaginarse a las criaturas inconscientes como físicamente diferentes de las conscientes; por ejemplo, exhiben una conducta deteriorada. Pero la verdadera cuestión no es la plausibilidad de que los zombis puedan existir en nuestro mundo, ni tampoco la naturalidad de la idea de una réplica zombi; el problema es si la noción de zombi es conceptualmente coherente. La mera inteligibilidad de la noción es suficiente para establecer la conclusión.

Una argumentación en favor de una posibilidad lógica no es del todo directa. ¿Cómo, por ejemplo, podríamos argumentar que un uniciclo de dos kilómetros de altura es lógicamente posible? Tan sólo parece obvio. Aunque no existe algo así en el mundo real, la descripción parece coherente. Si alguien objeta que no es lógicamente posible —tan sólo lo parece— hay poco que podamos hacer, excepto repetir la descripción y afirmar su obvia coherencia. Parece bastante evidente que no hay ninguna contradicción oculta en la descripción.

Confieso que la posibilidad lógica de los zombis me parece igualmente obvia. Un zombi es solamente alguien físicamente idéntico a mí, pero que no tiene experiencia consciente, todo está oscuro adentro. Aunque es probable que esto sea empíricamente imposible, pareciera que estamos describiendo una situación coherente; no puedo discernir ninguna contradicción en la descripción. En cierta forma una aserción de esta posibilidad lógica se reduce a una intuición primitiva, pero no más que en el caso del uniciclo. Casi todos, me parece, son capaces de concebir esta posibilidad. Algunos podrían

sentirse impulsados a negar la posibilidad para lograr que alguna teoría resulte bien, pero la justificación de tales teorías debería ser contingente a la cuestión de posibilidad, y no al revés.

En general, la carga de la prueba recae sobre aquellos que afirman que una determinada descripción es lógicamente *imposible*. Si alguien cree de verdad que un uniciclo de dos kilómetros es lógicamente imposible, debe darnos alguna idea de dónde se encuentra la contradicción, explícita o implícita. Si no puede señalar algo en las intensiones de los conceptos "dos kilómetros de alto" o "uniciclo" que pudiese llevar a una contradicción, entonces su postura no será muy convincente. Por otro lado, no resulta más convincente tampoco realizar un análisis obviamente falso de las nociones en cuestión; afirmar, por ejemplo, que para que algo califique de uniciclo debe ser más bajo que la Estatua de la Libertad. Si ningún análisis razonable de los términos en cuestión señala una contradicción o hace, por lo menos, que la existencia de una contradicción sea plausible, entonces existe un suposición natural en favor de la posibilidad lógica.

Dicho esto, hay algunas cosas positivas que los defensores de la posibilidad lógica pueden hacer para fundamentar su tesis. Pueden formular diversos argumentos indirectos, que recurren a lo que sabemos acerca de los fenómenos en cuestión y al modo como pensamos acerca de casos hipotéticos que involucran esos fenómenos, para establecer que la posibilidad lógica obvia es realmente una posibilidad lógica y es verdaderamente obvia. Podríamos crear una fantasía acerca de una persona ordinaria que está andando en un uniciclo, cuando repentinamente todo el sistema se expande mil veces. O podríamos describir una serie de uniciclos, cada uno más grande que el anterior. En cierto sentido, estos son todos recursos a la intuición, y un opositor que desee negar la posibilidad puede en cada caso afirmar que nuestras intuiciones nos llevaron a un error, pero la propia obviedad de lo que estamos describiendo trabaja a nuestro favor y ayuda a desplazar la carga de la prueba más hacia el otro lado.

Por ejemplo, podemos apoyar en forma indirecta la aseveración de que los zombis son lógicamente posibles considerando *realizaciones no estándar* de mi organización funcional.[2] Esta —o sea, el patrón de la organización causal incorporado en los mecanismos responsables de la producción de mi conducta— puede, en principio, realizarse en toda clase de modos extraños. Para usar un ejemplo común (Block, 1978) las personas de un país grande como China podrían organizarse para realizar una organización causal isomórfica a la de mi cerebro; cada persona simularía la conducta de una sola neurona, y poseería radioenlaces que corresponden a las sinapsis. La población

podría controlar un caparazón vacío de un cuerpo robot equipado con transductores sensoriales y efectores motores.

Para muchas personas no es plausible que una configuración como esta pueda dar origen a la experiencia consciente, esto es que, de algún modo, pueda surgir una "mente grupal" a partir del sistema total. No me interesa aquí la cuestión de si *en los hechos* surgiría o no la experiencia consciente aunque sospecho que sí lo haría, como argumentaré en el capítulo 7. Lo único que aquí me importa es que la idea de que un sistema de este tipo pueda carecer de experiencia consciente es *coherente*. Estamos expresando una posibilidad significativa, y es una pregunta abierta si la conciencia surge o no. Podemos señalar algo similar si consideramos un isomorfo mío de silicio, que está organizado como yo pero que tiene chips de silicio donde yo tengo neuronas. Si *en los hechos* este isomorfo sería consciente es una cuestión controversial, pero a la mayoría de las personas les parece que aquellos que lo niegan expresan una posibilidad coherente. A partir de casos como estos se deduce que la existencia de mi experiencia consciente no está lógicamente implicada por los hechos acerca de mi organización funcional.

Pero, dado que es conceptualmente coherente que la configuración de mente grupal o mi isomorfo de silicio puedan carecer de experiencia consciente, se deduce que mi gemelo zombi es una posibilidad igualmente coherente. Esto se debe a que es evidente que una implicación *conceptual* desde la bioquímica a la conciencia no tiene un mayor grado de posibilidad que una implicación desde el silicio o desde un grupo de homúnculos. Si el isomorfo de silicio carente de experiencia consciente es concebible, sólo debemos sustituir el silicio por neuronas en nuestra descripción, dejar la organización funcional constante, y tenemos entonces a mi gemelo zombi. Nada en esta sustitución nos fuerza a introducir la experiencia en la descripción; estas diferencias de implementación simplemente no son la clase de cosas que podrían ser conceptualmente relevantes para la experiencia. De este modo, la conciencia no superviene lógicamente a lo físico.

El argumento en favor de los zombis puede hacerse sin recurrir a estas realizaciones no estándar, pero ellas tienen un valor heurístico debido a que eliminan una fuente de confusión conceptual. Para algunas personas, las intuiciones acerca de la posibilidad lógica de una réplica física inconsciente no son muy claras al principio; esto quizá se deba a que la familiar coocurrencia de la bioquímica y la conciencia puede llevarnos a suponer una conexión conceptual. La consideración de los casos menos familiares elimina estas correlaciones empíricas del cuadro y, por lo tanto, hace que los juicios sobre la posibilidad lógica sean

más directos.[3] Pero una vez que se acepta que estas réplicas funcionales no conscientes son lógicamente posibles, la conclusión correspondiente acerca de una réplica física no puede evitarse.

Algunos podrían pensar que los argumentos basados en la conceptibilidad no son fiables. Por ejemplo, a veces se objeta que no podemos realmente imaginarnos en detalle los muchos miles de millones de neuronas en el cerebro humano. Por supuesto, esto es verdad; pero no tenemos necesidad de imaginarnos cada una de las neuronas para definir la cuestión. La mera complejidad entre neuronas no podría conceptualmente implicar a la conciencia; si toda esa estructura neuronal debe ser relevante para la conciencia, debe serlo *en virtud* de algunas propiedades de nivel superior que hace posibles. De este modo, es suficiente imaginar el sistema en un nivel no muy detallado y asegurarse que lo concebimos con mecanismos apropiadamente sofisticados de percepción, categorización, acceso de gran ancho de banda a los contenidos de información, capacidad de informar y otros similares. No importa lo sofisticados que imaginemos que son estos mecanismos, el escenario zombi seguirá siendo tan coherente como siempre. Quizás, un opositor podría sostener que todos los detalles neuronales no imaginados son conceptualmente relevantes de algún modo independiente de su contribución al funcionamiento sofisticado; pero, entonces, nos debe una explicación de cual podría ser ese modo, y no parece haber ninguno disponible. Esos detalles de implementación sencillamente se encuentran en el nivel erróneo para ser conceptualmente relevantes a la conciencia.

También se dice a veces que la conceptibilidad es una guía imperfecta de la posibilidad. El modo principal en que la conceptibilidad y la posibilidad pueden separarse está vinculado con el fenómeno de la necesidad *a posteriori*: por ejemplo, la hipótesis de que el agua no sea $H_2O$ parece conceptualmente coherente, pero es plausible que el agua sea $H_2O$ en todos los mundos posibles. Sin embargo, la necesidad *a posteriori* es irrelevante para las preocupaciones de este capítulo. Como vimos en el capítulo anterior, las conexiones explicativas se basan en implicaciones *a priori* de los hechos físicos con los hechos de alto nivel. La clase relevante de posibilidad debe ser evaluada usando las intensiones primarias de los términos involucrados, en lugar de las intensiones secundarias que son relevantes a la necesidad *a posteriori*. De esta forma, si un mundo zombi sólo fuera concebible en el sentido en que es concebible que el agua no sea $H_2O$, eso basta para establecer que la conciencia no puede explicarse reductivamente.

Dejando de lado estas consideraciones, los argumentos de conceptibilidad pueden descarriarse principalmente por una sutil confu-

sión conceptual: si somos insuficientemente reflexivos podemos pasar por alto una incoherencia en una supuesta posibilidad al tomar una situación concebida y *describirla erróneamente*. Por ejemplo, es posible pensar que podemos concebir una situación en la cual el último teorema de Fermat sea falso, imaginando una situación en la que matemáticos líderes declaran que encontraron un contraejemplo. Pero si el teorema es en realidad verdadero, esta situación está siendo descripta erróneamente: es en realidad un escenario en el cual el último teorema de Fermat es verdadero y algunos matemáticos cometen un error. Es importante, sin embargo, que esta clase de error siempre se encuentra en el dominio *a priori*, ya que surge de la aplicación incorrecta de las intensiones primarias de nuestros conceptos a una situación concebida. Una nivel de reflexión apropiado revelará que los conceptos están siendo aplicados incorrectamente, y que la aseveración de la posibilidad lógica no está justificada.

De este modo, el único camino disponible aquí para un opositor es afirmar que al describir el mundo zombi como un mundo zombi estamos aplicando erróneamente los conceptos y que, de hecho, hay una contradicción conceptual escondida en la descripción. Quizá, si pensásemos en ello de un modo suficientemente claro nos daríamos cuenta de que al imaginarnos un mundo físicamente idéntico estamos, por consiguiente, *automáticamente* imaginando un mundo en el cual existe la experiencia consciente. Pero, entonces, le corresponde al oponente darnos alguna idea de dónde podría encontrarse la contradicción en una descripción aparentemente bastante coherente. Si no puede mostrarse que exista alguna incoherencia interna, entonces tenemos una defensa muy sólida de que el mundo zombi es lógicamente posible.

Como antes, yo no puedo detectar ninguna incoherencia interna; tengo una clara imagen de lo que estoy concibiendo cuando concibo un zombi. Sin embargo, a algunas personas les resulta difícil evaluar los argumentos de conceptibilidad, en particular cuando están involucradas ideas extrañas como esta. Es, por lo tanto, afortunado que cada punto sostenido utilizando a los zombis pueda sostenerse también de otros modos, por ejemplo, considerando la epistemología y el análisis. Para muchos, los argumentos de estas últimas clases (como los argumentos 3 a 5 que siguen) son más directos y por lo tanto constituyen una base más sólida para el argumento en contra de la superveniencia lógica. Pero los zombis proporcionan, al menos, una vívida ilustración de importantes cuestiones afines.

### Argumento 2: El espectro invertido

Para ofrecer un argumento de conceptibilidad en contra de la superveniencia lógica, no es estrictamente necesario establecer la posibilidad lógica de los zombis o de un mundo de zombis. Es suficiente establecer la posibilidad lógica de un mundo físicamente idéntico al nuestro en el cual los hechos acerca de la experiencia consciente son meramente *diferentes* de los hechos en nuestro mun-do; no es necesario que la experiencia consciente esté totalmente ausente. Si algún hecho positivo acerca de la experiencia en nuestro mundo no es válido en un mundo físicamente idéntico, entonces la conciencia no es lógicamente superveniente.

Por lo tanto, es suficiente advertir que podemos imaginar de forma coherente un mundo físicamente idéntico en el cual las experiencias conscientes están *invertidas*, o (en el nivel local) imaginar un ser físicamente idéntico a mí pero con experiencias conscientes invertidas. Podríamos imaginar, por ejemplo, que cuando tengo una experiencia de color rojo, mi gemelo invertido tiene una experiencia de azul, y viceversa. Por supuesto, él llamará a sus experiencias de azul "rojo", pero eso es irrelevante. Lo importante es que la experiencia que él tiene de las cosas que ambos llamamos "rojas" —la sangre, las bombas de incendio, etc.— es del mismo tipo que la experiencia que yo tengo de las cosas que ambos llamamos "azules", como el mar y el cielo.

El resto de sus experiencias de color están sistemáticamente invertidas respecto de las mías, de forma de que sean coherentes con la inversión rojo-azul. Quizás el mejor modo de imaginar como esto ocurre con las experiencias de color humanas sea imaginar que dos de los ejes de nuestro espacio tridimensional del color están intercambiados: el eje rojo-verde se aplica en el eje amarillo-azul, y viceversa.[4] Para lograr una inversión en el mundo real, posiblemente necesitaríamos reconectar los procesos neuronales de un modo apropiado, pero como posibilidad lógica, parece totalmente coherente que las experiencias puedan invertirse mientras que la estructura física se duplica exactamente. Nada en la neurofisiología dicta que un tipo de procesamiento deba estar acompañado por experiencias de rojo y no por experiencias de amarillo.

A veces se objeta (Harrison, 1973; Hardin, 1987) que el espacio de color humano es asimétrico de un modo que impide una inversión de este tipo. Por ejemplo, ciertos colores tienen una calidez o frialdad asociados, y estas propiedades parecen estar directamente asociadas con diferentes papeles funcionales (por ejemplo, la calidez se percibe como "positiva", mientras que la frialdad se percibe como

"negativa"). Si un color cálido y un color frío se intercambian, entonces la sensación fenoménica "cálido" estaría disociada del papel funcional "cálido": una experiencia "fría" de verde se informaría como positiva en lugar de negativa, etc. De un modo similar, parece haber más matices discriminables de rojo que de amarillo, de forma que el intercambio de las experiencias de rojo con las experiencias de amarillo llevaría a la extraña situación en la cual un sujeto podría discriminar funcionalmente más matices de amarillo que los que son distinguibles fenomenológicamente. Quizás existan suficientes asimetrías en el espacio del color como para que cualquier inversión lleve a una extraña disociación de las sensaciones fenoménicas del papel funcional "apropiado".

Hay tres cosas que podemos decir en respuesta a esto. Primero, parece no haber nada *incoherente* en la noción de esta disociación (por ejemplo, la fenomenología fría con reacciones cálidas), aunque debe admitirse que es una idea algo extraña.[5] Segundo, en lugar de aplicar el rojo precisamente en el azul y viceversa, podemos imaginar que estos se aplican en colores levemente diferentes. Por ejemplo, el rojo podría aplicarse en una versión "cálida" del azul (como Levine [1991] sugiere), o incluso en un color que no esté en nuestro espacio de colores. En el caso rojo-amarillo, podemos imaginarnos que el rojo se aplica en un rango extendido de experiencias de amarillo, en el que hay disponible una mayor discriminación. No hay ninguna razón para que los escenarios de inversión de espectro *deban* involucrar colores extraídos del espacio usual del color. Tercero, quizá la respuesta más convincente sea argumentar (con Shoemaker [1982]) que aunque nuestro propio espacio del color es asimétrico, *podría* haber criaturas cuyo espacio de color sea simétrico. Por ejemplo, es probable que haya una criatura naturalmente posible que vea (y experimente) precisamente dos colores A y B, que corresponden a rangos distintos y bien separados de longitudes de ondas lumínicas, y cuya distinción agota la estructura del espacio del color. Parece totalmente coherente imaginarse dos de estas criaturas físicamente idénticas, pero cuyas experiencias de A y B están invertidas. Eso es suficiente para plantear la cuestión.

Incluso muchos materialistas reductivos (por ejemplo, Shoemaker, 1982) aceptaron que es coherente que nuestras experiencias de color puedan estar invertidas mientras que nuestra organización funcional se mantiene constante. Es permisible que un sistema con propiedades neurofisiológicas subyacentes diferentes, o con algo como silicio en lugar de neurobiología, pueda tener experiencias de color diferentes. Pero una vez que se acepta esto, le sigue automáticamente que la inversión de las experiencias en una réplica física es

al menos conceptualmente coherente. Las propiedades neurofisiológicas extra que están restringidas en un caso semejante nuevamente no son el tipo de cosas que podría determinar lógicamente la naturaleza de la experiencia. Aunque exista algún tipo de identificación *a posteriori* entre ciertas estructuras neurofisiológicas y ciertas experiencias (como cree Shoemaker), debemos aceptar que un patrón diferente de asociaciones es concebible, en el sentido de conceptibilidad que es relevante para la explicación reductiva.

Aunque la posibilidad del espectro invertido y la posibilidad de los zombis establecen que la conciencia no superviene lógicamente, la primera lleva a una conclusión estrictamente más débil que la segunda. Es concebible que alguien pueda sostener que el espectro invertido es lógicamente posible pero no los zombis. Si este fuera el caso, entonces la *existencia* de la conciencia podría explicarse reductivamente, pero el *carácter* específico de experiencias conscientes particulares no podría serlo.

### Argumento 3: De la asimetría epistémica

Como vimos antes, la conciencia es una característica sorprendente del universo. Nuestras razones para creer en la conciencia se derivan exclusivamente de nuestro propia experiencia de ella. Aun si conociésemos cada detalle acerca de la física del universo —la configuración, causalidad y la evolución de todos los campos y partículas en el continuo espaciotemporal— *esa* información no nos llevaría a postular la existencia de la experiencia consciente. Mi conocimiento de la conciencia proviene, en primera lugar, de mi propia persona, no de alguna otra observación externa. Es mi experiencia de primera persona de la conciencia lo que me fuerza a enfrentar el problema.

De todos los hechos de bajo nivel acerca de las configuraciones físicas y la causalidad, podemos en principio derivar toda clase de hechos de alto nivel acerca de los sistemas macroscópicos, su organización y la causalidad entre ellos. Podríamos determinar todos los hechos acerca del funcionamiento biológico, la conducta humana y los mecanismos cerebrales que la causan. Sin embargo, nada en esta vasta historia causal llevaría a quien no la haya experimentado directamente a creer que debería haber *una conciencia*. La misma podría incluso parecer irrazonable; casi mística quizás.

Es verdad que los hechos físicos acerca del mundo podrían proporcionar alguna evidencia indirecta en favor de la existencia de la conciencia. Por ejemplo, a partir de estos hechos podríamos averiguar que hay muchos organismos que *afirman* ser conscientes y dicen tener misteriosas experiencias subjetivas. Con todo, esta eviden-

cia sería bastante inconcluyente, y podría ser muy natural extraer una conclusión eliminativa: que de hecho no hay ninguna *experiencia* presente en estas criaturas, sólo mucha charla.

El eliminativismo acerca de la experiencia consciente es una posición irrazonable *sólo* debido a nuestro conocimiento de ella. Si no fuese por este conocimiento directo, la conciencia podría seguir el camino del espíritu vital. Para decirlo de otro modo, existe una *asimetría epistémica* en nuestro conocimiento de la conciencia que no existe en nuestro conocimiento de otros fenómenos.[6] Nuestro conocimiento de que la experiencia consciente existe se deriva principalmente de nuestro propio caso, y la evidencia externa tienen un papel cuanto más secundario.

También podemos defender la cuestión señalando la existencia del problema de las otras mentes. Aun cuando conozcamos todas las características físicas de otras criaturas, no *sabremos* con seguridad que son conscientes, o cuáles son sus experiencias (aunque podemos tener buenas razones para creer que lo son). Es sorprendente que no haya ningún problema con las "otras vidas" o las "otras economías" o las "otras alturas". No existe ninguna asimetría epistémica en estos casos, precisamente porque estos fenómenos son lógicamente supervenientes a lo físico.

La asimetría epistémica en el conocimiento de la conciencia hace evidente que la conciencia no puede ser lógicamente superveniente. Si lo fuese, no existiría esta asimetría epistémica; una propiedad lógicamente superveniente puede detectarse fácilmente sobre la base de la evidencia externa, y el caso de primera persona no tiene ningún papel especial. Seguramente, existen algunas propiedades supervenientes —la memoria, quizá— que son más fácilmente detectables en el caso de primera persona. Pero esto es sólo una cuestión de la dificultad de la tarea. En principio, la presencia de la memoria es tan accesible desde la tercera persona como desde la primera. La asimetría epistémica asociada a la conciencia es mucho más fundamental, y nos dice que ninguna colección de hechos acerca de la compleja causalidad en los sistemas físicos puede llegar a totalizar un hecho acerca de la conciencia.

### Argumento 4: El argumento a partir del conocimiento

El argumento más vívido en contra de la superveniencia lógica de la conciencia fue sugerido por Jackson (1982), siguiendo argumentos relacionados de Nagel (1974) y otros. Imagínese que estamos viviendo en una era en que la neurociencia ha sido terminada, en la que sabemos todo lo que hay que saber acerca de los procesos físicos

dentro de nuestro cerebro que son responsables de la producción de nuestra conducta. María creció en una habitación blanca y negra y nunca vio ningún color excepto el negro, el blanco y matices de gris.[7] Ella es, sin embargo, una de las neurocientíficas líderes, especializada en la neurofisiología de la visión del color. Sabe todo lo que hay que saber acerca de los procesos neuronales involucrados en el procesamiento de la información visual, la física de los procesos ópticos y la composición física de los objetos en el ambiente. Pero no sabe cómo es ver el color rojo. Ninguna cantidad de razonamiento a partir de los hechos físicos solamente podrá darle ese conocimiento.

Se deduce que los hechos sobre la experiencia subjetiva de la visión del color no están implicados por los hechos físicos. Si lo estuviesen, María podría en principio llegar a saber cómo es ver el color rojo sobre la base de su conocimiento de los hechos físicos. Pero no puede. Quizá María pudiese llegar a conocer cómo es ver el color rojo por medio de algún método indirecto, como manipular el cerebro del modo apropiado. El punto, sin embargo, es que el conocimiento no surge del conocimiento físico solamente. En principio, el conocimiento de todos los hechos físicos le permitirán a María derivar todos los hechos acerca de las reacciones, habilidades y capacidades cognitivas de un sistema; pero todavía estará totalmente a oscuras acerca de su experiencia del color rojo.

Un modo relacionado de formular esta cuestión es considerar sistemas muy diferentes de nosotros mismos, quizá mucho más simples —como los murciélagos o los ratones— y advertir que los hechos físicos acerca de estos sistemas no nos dicen cómo son sus experiencias, si es que las tienen. (Nagel se concentra en este tipo de cuestión.) Una vez que tenemos todos los hechos físicos acerca de un ratón, la naturaleza de su experiencia consciente sigue siendo una *pregunta abierta*: es consistente con los hechos físicos relativos a un ratón que tenga experiencia consciente y es consistente con los hechos físicos que no la tenga. De los hechos físicos acerca de un murciélago, podemos averiguar *todos* los hechos sobre un murciélago, excepto los hechos sobre sus experiencias conscientes. Aun conociendo todos los hechos físicos, no sabemos cómo se siente ser un murciélago.

A lo largo de lineamientos similares, podemos considerar un ordenador, diseñado como un agente cognitivo simple (quizá tenga la inteligencia de un perro), pero similar a nosotros en ciertos aspectos, como su capacidad para la discriminación perceptual. En particular categoriza los estímulos de color de una manera muy similar a la nuestra, agrupando cosas que llamaríamos "rojas" bajo una categoría y cosas que llamaríamos "verdes" bajo otra. Aun si conociésemos cada detalle acerca de los circuitos del ordenador, persistirían algunas

preguntas: 1) ¿El ordenador experimenta algo cuando mira una rosa?; 2) Si así es, ¿experimenta la misma cualidad sensorial de color que nosotros cuando miramos una rosa, o alguna cualidad diferente? Estas son preguntas totalmente significativas, y conocer todos los hechos físicos no fuerza una respuesta más que otra. Los hechos físicos, por lo tanto, no implican lógicamente a los hechos sobre la experiencia consciente.

Jackson realizó su formulación como un argumento en contra del materialismo más que en contra de la explicación reductiva. Han habido muchas respuestas al argumento; las analizaré en el siguiente capítulo, donde el materialismo, más que la explicación reductiva, será lo que esté en cuestión. Pero, por ahora, será interesante notar que la mayoría de las objeciones al argumento en contra del materialismo *aceptaron* la cuestión que es relevante en el argumento en contra de la explicación reductiva: que el conocimiento acerca de cómo es experimentar el color rojo es conocimiento fáctico que no está implicado *a priori* por el conocimiento de los hechos físicos. El único modo como puede evitarse la conclusión es negar que el conocimiento de cómo es la experiencia del color rojo pueda darnos algún conocimiento acerca del *hecho*. Esta es la estrategia adoptada por Lewis (1990) y Nemirow (1990), quienes argumentan que todo lo que le falta a María es una *destreza*, como por ejemplo la capacidad de reconocer cosas rojas. Analizaré esta sugerencia en el próximo capítulo; aquí, simplemente hago notar que dado que parece evidente que cuando ella ve el color rojo por primera vez está *descubriendo* algo acerca del modo de ser del mundo, resulta evidente que el conocimiento que ella obtiene es conocimiento sobre un hecho.

### Argumento 5: A partir de la ausencia de análisis

Si los defensores de la explicación reductiva abrigan alguna esperanza de vencer los argumentos de más arriba, deberán darnos alguna idea de cómo la existencia de la conciencia *podría* estar implicada por hechos físicos. Aunque no es justo esperar que nos suministren todos los detalles, al menos necesitamos una concepción de qué forma *podría* tomar una implicación de este tipo. Pero, cualquier intento por demostrar una implicación semejante está condenado al fracaso. Para que la conciencia esté implicada por un conjunto de hechos físicos, necesitaríamos algún tipo de análisis de la noción de conciencia —el tipo de análisis cuya satisfacción los hechos físicos podrían implicar— y no parece haber ningún análisis de esta clase.

El único análisis de la conciencia que parece tan siquiera remotamente sostenible para estos propósitos es un análisis funcio-

nal. Basándose en este análisis, se vería que lo que significa que algo esté consciente es que debería desempeñar un cierto papel funcional. Por ejemplo, podríamos decir que todo lo que significa que un estado sea consciente es que sea verbalmente comunicable, o que sea el resultado de ciertas clases de discriminación perceptual, o que haga que cierta información esté disponible de un cierto modo para procesos posteriores, o lo que sea. Pero a primera vista, estos fallan miserablemente como análisis. Simplemente fracasan en decir qué significa que algo sea una experiencia consciente. Aunque los estados conscientes pueden desempeñar diversos papeles causales, no están *definidos* por ellos. Más bien, lo que los hace conscientes es que poseen una cierta sensación fenoménica, y esta sensación no es algo que pueda ser eliminado funcionalmente.

Para ver cuán insatisfactorios son estos análisis, nótese cómo trivializan el problema de explicar la conciencia. Repentinamente, todo lo que debemos hacer para explicar la conciencia es explicar nuestra capacidad para hacer ciertos informes verbales, o para realizar ciertos tipos de discriminación, o manifestar alguna otra capacidad. Pero, a primera vista, es totalmente concebible que podamos explicar todas estas cosas sin explicar ni un ápice acerca de la propia conciencia; esto es, sin explicar la *experiencia* que acompaña al informe o a la discriminación. Analizar la conciencia en términos de alguna noción funcional es cambiar de tema o definir eliminativamente el problema. De la misma manera podríamos definir "paz mundial" como "un emparedado de jamón". Lograr la paz mundial se vuelve mucho más fácil, pero es un logro hueco.

También se puede argumentar en contra de los análisis funcionales de la conciencia sobre bases más específicas. Por ejemplo, cualquier concepto analizado funcionalmente tendrá un grado de indeterminación semántica. ¿Un ratón tiene creencias? ¿Las bacterias aprenden? ¿Un virus computacional está vivo? La mejor respuesta a estas preguntas es, por lo general, en un sentido sí y en un sentido no. Todo depende de cómo tracemos las fronteras entre los conceptos y para cualquier concepto funcional de alto nivel las fronteras serán vagas. Pero, compárese: ¿Un ratón tiene experiencia consciente? ¿Y un virus? Estas no son cuestiones de estipulación. Hay algo que es como ser un ratón o no lo hay, y no depende de nosotros definir la experiencia del ratón en o fuera de existencia. Es probable que exista un continuo de la experiencia consciente desde muy tenue a muy rica; pero si algo tiene experiencia consciente, aunque sea débil, no podemos eliminarla por estipulación. Esta determinación no podría derivarse de ningún análisis funcional de los conceptos afines de la conciencia, ya que los conceptos funcionales en dicha vecindad son

todos algo vagos. Entonces, se deduce que la noción de conciencia no puede analizarse funcionalmente.

Otra objeción es que el análisis funcionalista funde la distinción importante, descripta en el capítulo 1, entre las nociones de percatación y conciencia. Es de suponer que si la conciencia debe ser analizada funcionalmente, lo será aproximadamente como analizamos en ese momento la percatación: en términos de una cierta accesibilidad de la información durante el procesamiento posterior y en términos del control de la conducta. La percatación es un concepto perfectamente bueno, pero es bastante distinto del concepto de la experiencia consciente. El tratamiento funcionalista funde las dos nociones de conciencia y percatación en una y, por lo tanto, no hace justicia a nuestro sistema conceptual.

Las alternativas al análisis funcional parecen ser aun peores. Es muy oscura la cuestión de la existencia de algún tipo de análisis apropiado para la explicación reductiva. La única alternativa podría ser un análisis estructural —quizá la conciencia pueda analizarse como algún tipo de estructura bioquímica— pero ese análisis sería aun más inadecuado. *Sea* o no la conciencia una estructura bioquímica, eso no es lo que "conciencia" *significa*. Analizar la conciencia de esta forma nuevamente trivializa el problema explicativo al cambiar de tema. El concepto de conciencia parece ser irreducible, ya que sólo parece caracterizable en términos de conceptos que involucran a la conciencia.

Nótese que esto es bastante diferente del tipo de irreducibilidad que a veces se supone que ocurre en los conceptos de alto nivel en general. Hemos visto que muchas nociones de alto nivel no tienen definiciones precisas y ningún análisis manejable en términos de condiciones necesarias y suficientes. Con todo, como vimos en el último capítulo, estos conceptos tienen por lo menos un análisis aproximado que nos permitirá introducirnos en el campo de juego, aunque inevitablemente fracasará en hacer justicia a los detalles. Es fácil ver que propiedades como vida, aprendizaje, etc. pueden analizarse como propiedades funcionales, aunque especificar los detalles de exactamente *cuál* propiedad funcional es una cuestión difícil. Aun cuando estas propiedades carecen de definiciones funcionales precisas, son bastante compatibles, sin embargo, con la implicación por los hechos físicos.

Los problemas con la conciencia pertenecen a una liga diferente. Aquí, los supuestos análisis ni siquiera nos introducen en el campo de juego. De un modo mucho más severo, fracasan completamente en caracterizar lo que debe ser explicado. Ni siquiera existe la tentación de *intentar* agregar epiciclos a un supuesto análisis funcional de la

conciencia para hacerlo satisfactorio, como ocurre con análisis similares de la vida y el aprendizaje. La conciencia simplemente no debe ser caracterizada como una propiedad funcional en primer lugar. Lo mismo ocurre con los análisis de la conciencia como propiedad estructural o en otros términos reductivos. Por lo tanto, no hay forma de que una implicación de hechos físicos en la conciencia pueda despegar.

## 2. El fracaso de la explicación reductiva

La no superveniencia lógica a lo físico de la conciencia nos dice que ninguna explicación reductiva de la conciencia puede ser exitosa. Para cualquier concepción de los procesos físicos que supuestamente subyacen a la conciencia, siempre existirá una pregunta ulterior: ¿Por qué estos procesos están acompañados por la experiencia consciente? Para la mayoría de los otros fenómenos, esta pregunta puede responderse fácilmente: los hechos físicos acerca de esos procesos *implican* la existencia de los fenómenos. Para un fenómeno como la vida, por ejemplo, los hechos físicos implican la realización de ciertas funciones, y esto es todo lo que necesitamos explicar para explicar la vida. Pero una respuesta de esta clase no será suficiente para la conciencia.

La explicación física es muy apropiada para la explicación de la *estructura* y de la *función*. Las propiedades estructurales y las propiedades funcionales pueden ser directamente implicadas por una historia física de bajo nivel, por lo que son claramente aptas para la explicación reductiva. Y casi todos los fenómenos de alto nivel que debemos explicar se reducen, en última instancia, a estructura o función: piénsese en la explicación de los saltos de agua, los planetas, la digestión, la reproducción, el lenguaje. Pero la explicación de la conciencia no es sólo una cuestión de explicar la estructura y función. Una vez que explicamos toda la estructura física relacionada con el cerebro y explicamos cómo se realizan las diversas funciones cerebrales, queda una especie ulterior de explanandum: la propia conciencia. ¿Por qué debería toda esa estructura y función dar origen a la experiencia? La historia acerca de los procesos físicos no lo dice.

Podemos formular esto en términos de los experimentos mentales presentados anteriormente. Cualquier historia acerca de procesos físicos se aplica igualmente a mí y a mi gemelo zombi. Nada en esa historia dice por qué, en mi caso, surge la conciencia. De modo similar, cualquier historia acerca de procesos físicos se aplica igualmente a mi gemelo invertido que ve azul cuando yo veo rojo: nada en esa historia dice por qué mi experiencia es de una variedad y no de otra. El hecho de que es lógicamente posible que los hechos físicos puedan ser

iguales pero los hechos acerca de la conciencia sean diferentes nos muestra que, como Levine (1983) lo dijo, existe una *brecha explicativa* entre el nivel físico y la experiencia consciente.

Si esto es correcto, el hecho de que la conciencia acompañe a un proceso físico dado es un *hecho ulterior*, que no puede explicarse simplemente contando la historia acerca de los hechos físicos. En cierto sentido, este acompañamiento debe tomarse como una primitiva. Intentamos sistematizar y explicar estos hechos primitivos en términos de algún patrón simple subyacente, pero siempre quedará un elemento que es lógicamente independiente de la historia física. Quizá pudiésemos obtener algún tipo de explicación combinando los hechos físicos subyacentes con ciertos principios *puente* que vinculan los hechos físicos con la conciencia, pero esta explicación no será reductiva. La necesidad de principios puente explícitos nos muestra que la conciencia no está siendo explicada reductivamente, sino que está siendo explicada en sus propios términos.

Por supuesto, nada de lo que he dicho implica que los hechos físicos sean *irrelevantes* para la explicación de la conciencia. Todavía podemos esperar que las descripciones físicas desempeñen un papel significativo en una teoría de la conciencia, dando información acerca de la *base* física de la misma, por ejemplo, y quizá suministrando una correspondencia detallada entre diversos aspectos del procesamiento físico y aspectos de la experiencia consciente. Estas concepciones pueden ser especialmente útiles para ayudar a comprender la *estructura* de la conciencia: los patrones de similitud y diferencia entre experiencias, la estructura geométrica de campos fenoménicos, etc. Diré mucho más sobre estas y otras cosas que la explicación física puede contarnos acerca de la experiencia en un marco no reductivo en el capítulo 6. Pero una concepción física, solamente, no es suficiente.

En este punto, surge naturalmente un número de objeciones.

### Objeción 1: ¿Fijamos estándares demasiado elevados?

Algunos podrían argumentar que la explicación de *cualquier* fenómeno de alto nivel deberá postular "leyes puente" además de una descripción de bajo nivel, y que es sólo con la ayuda de estas leyes puente que pueden derivarse los detalles de los fenómenos de alto nivel. Sin embargo, como lo sugiere la discusión en el último capítulo (y se argumenta cuidadosamente en Horgan, 1978), en estos casos las leyes puente no son hechos ulteriores acerca del mundo. Más bien, los propios principios de conexión son lógicamente supervenientes a los hechos de bajo nivel. El caso extremo de un principio puente de este tipo es un condicional de superveniencia que, como ya vimos,

es por lo general una verdad conceptual. Otros principios puente más "localizados", tal como el vínculo entre el movimiento molecular y el calor, pueden al menos derivarse a partir de hechos físicos. Para la conciencia, en cambio, esos principios puente deben tomarse como primitivos.

Es interesante ver cómo una típica propiedad de alto nivel —como la vida, digamos— evade los argumentos formulados para el caso de la conciencia. Primero, es directamente inconcebible que pueda existir una réplica física de una criatura viviente que no esté viva. Quizá pudiese surgir un problema debido a propiedades dependientes del contexto (¿una réplica que se forma aleatoriamente en un pantano estaría viva?, ¿sería un ser humano?), pero fijar los hechos ambientales elimina incluso esa posibilidad. Segundo, no existe ninguna posibilidad de "vida invertida" análoga al del espectro invertido. Tercero, cuando conocemos todos los hechos físicos acerca de un organismo (y posiblemente acerca de su ambiente), tenemos suficiente material para conocer todos los hechos biológicos. Cuarto, no existe ninguna asimetría epistémica en el caso de la vida; los hechos acerca de la vida en los otros son tan accesibles, en principio, como los hechos acerca de la vida en nosotros mismos. Quinto, el concepto de vida es analizable en términos funcionales: estar vivo es aproximadamente poseer ciertas capacidades para adaptarse, reproducirse y metabolizar. En general, la mayor parte de los fenómenos de alto nivel se reducen a cuestiones de estructura física y función, y tenemos buenas razones para creer que las propiedades estructurales y funcionales son lógicamente supervenientes a lo físico.

### Objeción 2: ¿Un vitalista no podría haber dicho lo mismo acerca de la vida?

Todo esto no obstante, una reacción común al tipo de argumento que he formulado es responder que un vitalista podría haber dicho las mismas cosas acerca de la vida.[8] Por ejemplo, un vitalista podría haber afirmado que es lógicamente posible que una réplica física de mí podría no estar *viva*, para establecer que la vida no puede explicarse reductivamente. Y un vitalista podría haber argumentado que la vida es un hecho ulterior, no explicado por ninguna concepción de los hechos físicos. Sin embargo, el vitalista se habría *equivocado*. Por analogía, quien se oponga a la explicación reductiva de la conciencia ¿no podría él también equivocarse?

Creo que esta reacción comete un error al localizar la fuente de las objeciones vitalistas. El vitalismo estaba fundamentalmente impulsado por las dudas acerca de que los mecanismos físicos pudie-

sen realizar todas las *funciones* complejas asociadas con la vida: conducta adaptativa, reproducción, etc. En esa época, se sabía muy poco acerca de la enorme sofisticación de los mecanismos bioquímicos, de modo que este tipo de dudas era bastante comprensible. Pero, en estas mismas dudas se encuentra implícito el punto conceptual de que cuando se trata de explicar la vida, es la realización de las diversas funciones lo que debe explicarse. Es notable que a medida que gradualmente surgieron las explicaciones físicas de las funciones relevantes, las dudas vitalistas se disolvieron en su gran mayoría. Con la conciencia, en cambio, el problema persistirá aun cuando se expliquen las diversas funciones.

Ante una concepción física completa que muestre cómo los procesos físicos realizan las funciones relevantes, un vitalista razonable aceptaría que la vida ha sido explicada. Ni siquiera hay espacio *conceptual* para la realización de estas funciones sin vida. Podría ocurrir que algún vitalista ultra fanático pueda negar esto incluso, y afirmar que una concepción funcional de la vida deja algo afuera, quizás el espíritu vital. Pero la respuesta obvia es que, a diferencia de la experiencia, no tenemos ninguna razón independiente para creer en la existencia del espíritu vital. Si alguna vez hubo alguna razón para creer en él, era como construcción explicativa, "*Debe* existir ese tipo de cosa para poder hacer eso tan sorprendente". Pero, el espíritu vital como construcción explicativa puede eliminarse una vez que encontramos una mejor explicación de cómo se realizan las funciones. La experiencia consciente, en cambio, se nos impone como un explanandum y no puede eliminarse tan fácilmente.

Una razón por la que un vitalista podría pensar que algo quedó afuera de una explicación funcional de la vida sería precisamente que no hay nada en una concepción física que explique por qué hay algo que es como estar vivo. Quizás, algún elemento de la creencia en un "espíritu vital" estuvo vinculado a los fenómenos de la propia vida interior. Muchos percibieron un nexo entre los conceptos de vida y experiencia, e incluso hoy en día parece razonable decir que una de las cosas que deben explicarse acerca de la vida es el hecho de que muchas criaturas vivientes son conscientes. Pero la existencia de *este* tipo de duda vitalista no ofrece ninguna comodidad al defensor de la explicación reductiva de la conciencia, ya que es una duda que nunca ha sido resuelta.

## Objeción 3: ¿Es la conceptibilidad una guía de la posibilidad?

Los filósofos suelen sospechar de los argumentos que le asignan un papel clave a la conceptibilidad, y con frecuencia responden que la conceptibilidad no es suficiente para la posibilidad. Esta es una cuestión sutil que analicé anteriormente y volveré a hacerlo, pero aquí las sutilezas no son especialmente relevantes. Cuando se trata de cuestiones de *explicación*, es evidente que la conceptibilidad es fundamental. Si al reflexionar encontramos concebible que todos estos procesos físicos podrían ocurrir en ausencia de la conciencia, entonces ninguna explicación reductiva de la misma será satisfactoria: la pregunta ulterior de por qué *nosotros* existimos y no los zombis surgirá permanentemente. Aunque la conceptibilidad está ligada a los límites de la capacidad humana, la explicación también lo está de un modo similar.

Otro modo de formular esto es notar que la explicación reductiva de un fenómeno en términos de lo físico requiere una implicación *a priori* de los hechos físicos en los hechos de alto nivel relevantes (superveniencia lógica de acuerdo con la intensión primaria, tal como lo formulé antes). Si no existe una conexión de este tipo, entonces siempre podremos plantear la pregunta ulterior de por qué los procesos físicos dan origen a la conciencia. Hemos visto que en casi todos los dominios existe el tipo correcto de conexión, lo que hace posible la explicación reductiva; pero no parece ocurrir lo mismo para la experiencia consciente. Podemos poner en duda que los enfoques *ontológicos* como el materialismo giren en torno de estos vínculos *a priori* —analizaré esa cuestión en el próximo capítulo—, pero cuando se trata de la explicación reductiva, estos vínculos son cruciales.

## Objeción 4: ¿ No es esta una colección de intuiciones circulares?

Podría objetarse, además, que los argumentos que formulé consisten, en el fondo, en una colección de intuiciones. Ciertamente hay un sentido en el cual todos estos argumentos se basan en intuiciones, pero traté de aclarar lo natural y evidente que estas intuiciones son, y lo forzado que es negarlas. La intuición principal es que *hay algo que debe ser explicado*, algún fenómeno asociado con la experiencia de primera persona que plantea un problema no presentado por la observación de la cognición desde el punto de vista de tercera persona. Si se acepta la premisa de que la experiencia de primera persona nos impone algún explanandum que la observación de tercera persona no nos impone, la mayor parte de los argumentos

expuestos antes surgen naturalmente. Se deduce de inmediato, por ejemplo, que lo que requiere explicación no puede analizarse como el desempeño de algún papel funcional, porque este tipo de fenómenos se nos revela por la observación de tercera persona y es mucho más directo.

La "intuición" aquí actuante es la misma *raison d'être* del problema de la conciencia. El único modo consistente de eludir las intuiciones es directamente negar el problema y el fenómeno. Siempre podemos, al menos cuando hablamos "filosóficamente", negar por completo las intuiciones y negar que haya algo (aparte de la realización de diversas funciones) que requiera explicación. En cambio, si se toma a la conciencia seriamente, las conclusiones acerca de las que estoy argumentando deberán deducirse.

### Objeción 5: ¿No deben todas las explicaciones detenerse en algún lugar?

Una objeción final es que ninguna explicación nos da algo por nada: toda explicación tiene que detenerse en algún lugar. Al explicar el movimiento de los planetas, por ejemplo, damos por sentadas las leyes de la gravedad y la existencia de la masa. ¿No podría ocurrir aquí que simplemente deberíamos dar algo por sentado? Tengo una cierta simpatía por este argumento; creo que debemos dar algo por sentado para explicar la conciencia. Pero, al hacerlo así, inevitablemente nos movemos más allá de una explicación *reductiva*. Ciertamente, este tipo de analogía apoya la posición no reductiva que defiendo. Damos por sentadas las leyes de la física porque son leyes *fundamentales*. Si damos por supuesto un vínculo entre los procesos físicos y la experiencia consciente, esto sugiere que el vínculo también debería ser considerado fundamental. Volveré a este punto en el próximo capítulo.

## 3. La modelización cognitiva

En este y los apartados que siguen, ilustraré el fracaso de la explicación reductiva mediante una crítica de un número de concepciones de la conciencia propuestas por investigadores en diversas disciplinas. No todas estas propuestas han sido formuladas como explicaciones reductivas de la experiencia consciente, aunque por lo general fueron interpretadas de ese modo; pero de cualquier forma, es instructivo determinar exactamente qué pueden lograr y qué no estas concepciones. Durante la marcha, será interesante advertir las variables actitudes de estos investigadores hacia las preguntas difíciles acerca de la experiencia consciente.

Primero, consideraré las concepciones basadas en la *modelización cognitiva*. Esta funciona bien para la mayoría de los problemas en la ciencia cognitiva. Al exhibir un modelo de la dinámica causal involucrada en los procesos cognitivos, podemos explicar la causalidad de la conducta en un agente cognitivo. Esto proporciona un tipo valioso de explicación para fenómenos psicológicos como el aprendizaje, la memoria, la percepción, el control de la acción, la atención, la categorización, la conducta lingüística, etc. Si tenemos un modelo que capture la dinámica causal de alguien que está aprendiendo, por ejemplo, se deduce que cualquier cosa que instancie esa dinámica en el ambiente correcto estará aprendiendo. A partir del modelo podemos ver cómo se realizan ciertas funciones, y esto es todo lo que tenemos que aclarar para explicar el aprendizaje. Pero esto es insuficiente para explicar la conciencia. Para cualquier modelo que exhibamos, la pregunta ulterior de por qué su realización debería estar acompañada por la conciencia sigue vigente. Esta no es una cuestión que la descripción y el análisis del modelo solos puedan responder.

A veces se objeta que los supuestos modelos de la conciencia son inverificables, ya que no existe ninguna forma de verificar si instanciaciones de los mismos son o no conscientes. Este es un problema, pero existe otro aun más profundo. Incluso si tuviésemos (*per impossibile*) un "medidor de experiencia" que pudiese escudriñar y decirnos si una instanciación es consciente, esto sólo establecería una correlación. Sabríamos que cada vez que el modelo es instanciado, la conciencia lo acompaña. Pero no explicaría la conciencia de la manera como este tipo de modelos explica otros fenómenos mentales.

Esta clase de modelos puede ciertamente explicar la "conciencia" en los sentidos psicológicos considerados, en los que esta se interpreta como una especie de capacidad cognitiva o funcional. Muchos "modelos de la conciencia" existentes pueden ser muy generosamente interpretados bajo esta luz. Podemos considerar que proporcionan explicaciones de la informatividad, de la atención o de las capacidades introspectivas, etc. Ninguno de ellos, sin embargo, nos da algo que se acerque a una explicación de por qué estos procesos deberían estar acompañados por la experiencia consciente. Algunos ejemplos ilustrarán la cuestión.

El primer ejemplo es el modelo cognitivo presentado por Bernard Baars (1988), como parte de un tratamiento extenso de la conciencia desde el punto de vista de la psicología cognitiva. Baars aporta todo tipo de evidencias experimentales para fundamentar su tesis principal: la conciencia es un tipo de *espacio de trabajo global* en un sistema distribuido de procesadores inteligentes de información.

Cuando los procesadores acceden al espacio de trabajo global, transmiten un mensaje a todo el sistema, como si lo hubiesen escrito en un pizarrón. El contenido del espacio de trabajo global es el contenido de la conciencia.

Baars utiliza este modelo para explicar un número notable de propiedades del procesamiento humano. El modelo proporciona un marco muy sugerente para explicar el acceso del sujeto a la información, y su papel en la atención, la informatividad, el control voluntario e incluso el desarrollo de un autoconcepto. El marco del espacio de trabajo global es, por lo tanto, apropiado para explicar la conciencia en el conjunto completo de sus sentidos psicológicos. Se ofrece por lo menos una teoría general de la *percatación*.

Pero no puede encontrarse aquí ninguna explicación reductiva de la *experiencia*. La pregunta de por qué estos procesos deberían dar origen a la experiencia simplemente no se encara. Podríamos suponer que de acuerdo con la teoría, el contenido de la experiencia es precisamente el contenido del espacio de trabajo. Pero aun si esto es así, nada interno a la teoría *explica* por qué la información dentro del espacio de trabajo global es experimentada. Lo más que puede hacer es decir que la información es experimentada porque es *globalmente accesible*. Pero, entonces, surge la pregunta en una forma diferente: ¿Por qué la accesibilidad global debería dar origen a la experiencia consciente? Esta pregunta puente no ha sido encarada en el trabajo de Baars.

Baars menciona este tipo de preocupación brevemente: "Un lector escéptico podría (...) preguntarse si es verdad que estamos describiendo la experiencia consciente o si, en cambio, sólo podemos tratar con fenómenos incidentales asociados con ella" (p. 27). Su respuesta es hacer notar que las teorías científicas tienden al menos a *aproximarse* a la "cosa en sí misma"; por ejemplo, la biología explica la herencia *misma*, y no sólo fenómenos asociados. Pero, esto es simplemente ignorar los modos en los que la conciencia es categorialmente diferente de estos fenómenos, como ya hemos visto. En el caso de la herencia, las diversas funciones son todo lo que hay que explicar. En el caso de la conciencia, existe un explanandum ulterior: la propia experiencia. La teoría de Baars puede considerarse, por lo tanto, un enfoque interesante de los procesos cognitivos que subyacen a la conciencia, uno que nos ofrece mucha comprensión indirecta de la misma, pero que no toca las cuestiones claves: ¿por qué existe la conciencia y cómo surge a partir del procesamiento cognitivo?)

Daniel Dennett formuló también un modelo cognitivo de la conciencia. De hecho, formuló al menos dos. El primero (Dennett 1978c) es un "diagrama de flujo", consistente en una descripción de

la circulación de la información entre diversos módulos (fig. 3.2). Central al modelo son 1) un módulo perceptual, 2) una memoria a corto plazo M, que recibe información del módulo perceptual, 3) un sistema de control que interactúa con el almacenamiento de la memoria mediante un proceso de preguntas y respuestas y que puede dirigir la atención hacia el contenido del módulo perceptual, y 4) una unidad de "relaciones públicas", que recibe comandos de actos de habla del sistema de control y los convierte en emisiones en el lenguaje público.

¿Qué es lo que este modelo podría explicar? Aunque está en una forma muy simplificada (como Dennett aceptaría), sería posible especificarlo para que proporcione una explicación de la *informatividad*; esto es, de nuestra capacidad para informar sobre el contenido de nuestros estados internos. También proporciona el esquema de

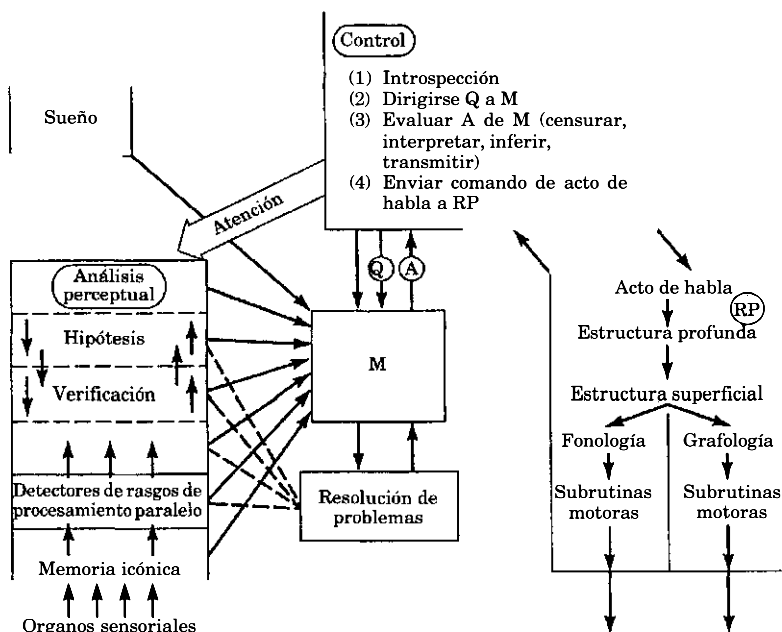



*Figura 3.2.* Modelo cognitivo de la conciencia de Dennett. (Redibujado de la fig. 9.1, p. 155, de Daniel C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, The MIT Press. Copyright © 1987 por Bradford Books, Publishers. Con autorización de The MIT Press.)

una explicación de nuestra capacidad de aportar información perceptual al control de la conducta, de hacer introspección sobre nuestros estados internos, etc. Pero no nos dice nada de por qué debería haber algo que es como ser un sistema que realiza estos procesos.

En *Consciousness Explained* (1991), Dennett formula una concepción más sofisticada que se basa en buena parte del trabajo reciente de la ciencia cognitiva. El modelo que aquí se propone es esencialmente un modelo de "pandemonio", consistente en muchos pequeños agentes que compiten por la atención, donde el agente que grita más fuerte desempeña el papel principal en la dirección del procesamiento posterior. En este modelo no hay ningún "cuartel general" de control, sino múltiples canales que ejercen influencia simultáneamente. Dennett complementa esta concepción recurriendo a la neurociencia, la biología evolutiva, los modelos conexionistas y los sistemas de producción en inteligencia artificial.

A pesar de la complejidad, este modelo está orientado principalmente a los mismos fenómenos que el anterior. Si fuera exitoso, proporcionaría una explicación de la informatividad y, más en general, de la influencia de los diversos tipos de información sobre el control de la conducta. El modelo citado da una explicación potencial del foco de atención. También presenta una concepción provocativa de algunas de nuestras capacidades cognitivas, pero no va más allá del modelo previo en decirnos por qué debería haber una experiencia consciente afín con estas capacidades.

A diferencia de la mayoría de los autores que formulan modelos cognitivos, Dennett sostiene explícitamente que los suyos son del tipo que podría explicar todo lo que en la experiencia necesita explicación. En particular, él piensa que para explicar la conciencia, sólo se necesita explicar fenómenos funcionales como la informatividad y el control; cualquier fenómeno que evidentemente se omita es una quimera. A veces, parece adoptar como premisa básica el que una vez que explicamos las diversas funciones, explicamos todo (véase, por ejemplo, Dennett, 1993a, p. 210), pero ocasionalmente formula argumentos que apoyan esto, algunos de los cuales consideraremos más adelante.[9]

Podría hacerse el mismo tipo de crítica a los enfoques de modelos cognitivos de la conciencia de Churchland (1995), Johnson-Laird (1988), Shallice (1972, 1988a, 1988b), y muchos otros. Todos presentan interesantes concepciones acerca de la realización de las funciones cognitivas, pero ninguna toca las preguntas realmente difíciles.

## 4. La explicación neurobiológica

Los enfoques neurobiológicos de la conciencia han adquirido popularidad en tiempos recientes. Como los modelos cognitivos, tienen mucho que ofrecer en la explicación de fenómenos psicológicos, tales como las variedades de la percatación. También nos pueden decir algo acerca de los procesos cerebrales que están *correlacionados* con la conciencia. Pero ninguna de estas concepciones explica la correlación: no nos dicen nada en absoluto de por qué los procesos cerebrales deberían dar origen a la experiencia. Desde el punto de vista de la neurociencia, la correlación es simplemente un hecho primitivo.

Desde un punto de vista metodológico, no es obvio cómo podemos comenzar a desarrollar una teoría neurocientífica. ¿Cómo se realizan los experimentos que detectan una correlación entre algún proceso neuronal y la conciencia? Lo que ocurre por lo general es que los teóricos se basan implícitamente en algún criterio psicológico de la conciencia, como el foco de atención, el control de la conducta, o más frecuentemente la capacidad para hacer informes verbales acerca de un estado interno. Luego advierten que alguna propiedad neurofisiológica está instanciada cuando estos criterios están presentes, y despega entonces una teoría de la conciencia.

El propio hecho de que se utilicen criterios indirectos de este tipo, sin embargo, hace evidente que no aportan ninguna explicación reductiva de la conciencia. Cuanto más, una concepción neurofisiológica podría ser capaz de explicar por qué se instancia la propiedad psicológica relevante. La pregunta de por qué la propiedad psicológica en cuestión debería estar acompañada por la experiencia consciente queda sin respuesta. Debido a que estas teorías se apoyan en el supuesto de un vínculo entre las propiedades psicológicas y la experiencia consciente, es evidente que no hacen nada para explicar ese vínculo. Esto se ve claramente mediante el examen de algunas concepciones neurocientíficas recientes de la conciencia.

Ultimamente concitaron mucha atención en la neurociencia ciertas oscilaciones de 40 hertz presentes en la corteza visual y otros sitios. Francis Crick y Christof Koch (1990) formularon la hipótesis de que este tipo de oscilaciones podría ser la característica neuronal fundamental responsable de la experiencia consciente, y propulsaron el desarrollo de una teoría neurobiológica según estas líneas.[10]

¿Por qué oscilaciones de 40 hertz? Principalmente porque la evidencia sugiere que estas oscilaciones tienen un papel importante en la *asociación* de diversos tipos de información en un todo unificado. Dos clases diferentes de información sobre una escena —la forma y

localización de un objeto, por ejemplo—pueden recibir representaciones relativamente independientes, pero esta teoría sugiere que las diferentes representaciones neuronales pueden tener una frecuencia y fase común en sus oscilaciones, lo que permitiría que la información se vincule mediante procesos posteriores y se almacene en la memoria de trabajo. De esta forma, todo tipo de información dispar podría ser integrada en el "contenido de la conciencia".

Una teoría de esta clase podría proporcionar una comprensión neurobiológica de la asociación y la memoria de trabajo, y quizás, luego, podría ser elaborada en una concepción de cómo se utiliza la información de un modo integrado en el control de la conducta. Pero la pregunta sigue en pie: ¿Por qué estas oscilaciones deberían estar acompañadas por la experiencia consciente? La teoría proporciona una respuesta parcial: porque estas oscilaciones son responsables de la asociación. Pero la pregunta de por qué la propia asociación debería estar acompañada por la experiencia no se encara. La teoría se apoya en el supuesto de un vínculo entre la asociación y la conciencia, y por lo tanto no hace nada por explicarla.

Crick y Koch parecen simpatizar con el "gran" problema de la conciencia, y lo llaman el "principal problema que enfrenta el enfoque neuronal de la mente". Argumentan que los enfoques de nivel puramente cognitivo están condenados al fracaso, y que se requieren teorías de nivel neuronal. Pero no nos dan ninguna razón para creer que su teoría sea más apropiada que las teorías cognitivas en lo que respecta a responder las preguntas realmente difíciles. Ciertamente, no sostienen que su proyecto resuelva el problema de la experiencia. En una entrevista publicada, Koch es bastante claro acerca de las limitaciones del enfoque:

> Bueno, primero olvidémonos de los aspectos realmente difíciles, como los sentimientos subjetivos, porque podrían no tener una solución científica. Los estados subjetivos del juego, el dolor, el placer, ver el azul, oler una rosa, parecen representar un salto enorme entre el nivel materialista, de explicación de moléculas y neuronas, y el nivel subjetivo. Debemos concentrarnos en cosas que son más fáciles de estudiar, como la percatación visual. Usted está ahora hablando conmigo, pero no me está mirando, está mirando ese capuchino, y por lo tanto se percata de él. Usted puede decir, "Es una taza y hay algún líquido adentro". Si yo se la alcanzo, usted moverá su brazo y la tomará; usted responderá de un modo significativo. Esto es lo que llamo percatación.[11]

Otra teoría neurofisiológica de la conciencia fue formulada por Gerald Edelman en *The Remembered Present* (1989) y otros libros y

artículos. El elemento central de su teoría involucra circuitos neuronales reentrantes mediante los cuales las señales perceptuales pueden ser categorizadas conceptualmente antes de que contribuyan a la memoria. La información perceptual y el estado interno interactúan de un modo sutil (como se describe en la fig. 3.3) para dar origen a la "conciencia primaria". Su modelo de la "conciencia de orden superior" aporta un nuevo elemento de memoria a través de una "inicialización [*bootstrapping*] semántica" que produce los conceptos de sí mismo, pasado y futuro. Todo esto está vinculado con la producción del lenguaje a través de las áreas de Broca y Wernicke.

Gran parte del trabajo de Edelman se dedica a la explicación de la percepción, la memoria, y el lenguaje, pero no a la explicación de la conciencia. Cuando habla de la conciencia, su exposición es vaga por lo general, pero pareciera que lo que en última instancia podría

SI MISMO
Sistemas
homeostáticos internos

NO SI MISMO
Señales del mundo
incluyendo la propiocepción

Tronco cerebral,
hipotálamo, centros
autonómicos

Corteza primaria y
secundaria de la visión,
audición, tacto, etcétera

Registro actual de
los estados y valores
internos

Categorización
perceptual actual

Inicialización
semántica

CONCIENCIA
PRIMARIA
Bucle reentrante
que conecta la
memoria de
valor-categoría
con la
categorización
perceptual actual

Correlación en
hipocampo,
amígdala, septum

Areas de
Broca y
Wernicke

Categorización
conceptual

CONCIENCIA DE
ORDEN SUPERIOR

Memoria especial
valor-categoría en
corteza frontal,
temporal, parietal

(Implica experiencia
previa y selección de
grupos neuronales)

*Figura 3.3.* Esquema de Edelman de la conciencia de orden superior. (Redibujado a partir de la figura 12.4, p. 132, de *Bright Air, Brilliant Fire,* de Gerald M. Edelman. Copyright © 1992 de BasicBooks, Inc. Con autorización de BasicBooks, una división de HarperCollins, Inc.)

explicarse mediante este tipo de modelo es la percatación perceptual —esto es, los efectos del procesamiento perceptual sobre procesos posteriores y sobre el control de la conducta— y aspectos de la autoconciencia, en especial el origen del concepto de sí mismo.

Edelman no ofrece ninguna concepción de cómo todo ese procesamiento debería dar origen a la experiencia consciente. Simplemente supone que hay una correlación. El es explícito sobre esto, y advierte que la experiencia fenoménica es el problema más difícil para una teoría de la conciencia, y que ninguna teoría física nos llevará todo el camino hasta los qualia:

> Esto sugiere un enfoque hacia el problema de los qualia. Como base de una teoría de la conciencia, es razonable *suponer* que, así como en nosotros, los qualia existen en otros seres humanos conscientes, ya sea que se los considere observadores científicos o sujetos .... Podemos considerar, entonces, a los seres humanos como el mejor referente canónico para el estudio de la conciencia. Esto está justificado por el hecho de que los informes subjetivos humanos (incluyendo aquellos sobre los qualia), las acciones y las estructuras cerebrales y funciones *pueden todos ponerse en correlación*. Luego de construir una teoría basada en el supuesto de que los qualia existen en los seres humanos, podemos entonces volver a examinar algunas de las propiedades de los qualia en base a estas correlaciones. Es nuestra capacidad de informar y correlacionar mientras individualmente experimentamos los qualia lo que abre la posibilidad de una investigación científica de la conciencia. (Edelman, 1992, p. 115)

Como antes, debido a que esta teoría se basa en el *supuesto* de correlación, es evidente que no ofrece una explicación reductiva de la experiencia. La mayor parte del tiempo, Edelman sólo sostiene que explica los procesos que subyacen a la experiencia consciente; no afirma estar explicando la experiencia.[12]

## 5. La apelación a la nueva física

A veces se sostiene que la clave de la explicación de la conciencia puede encontrarse en un nuevo tipo de teoría física. Quizás, al argumentar que la conciencia no está implicada por la física de nuestro mundo, hemos estado tácitamente suponiendo que la física de nuestro mundo es la física tal como la entendemos hoy en día, consistente de organizaciones de partículas y campos en el continuo espaciotemporal, que sufren complejos procesos causales y de evolución. Un opositor podría estar de acuerdo con nosotros en que nada en

*este* tipo de física entraña la existencia de la conciencia, pero argumentar también que podría existir un nuevo tipo de teoría física de la cual la conciencia surja como consecuencia.

No es fácil evaluar esta aseveración en ausencia de alguna propuesta detallada. Nos gustaría ver al menos un ejemplo de cómo *podría* ser esa física. No es necesario que un ejemplo de este tipo sea plausible a la luz de las teorías actuales, pero debería haber algún sentido en el cual se la pueda reconocer como una teoría física. La pregunta crucial es: ¿Cómo podría una teoría, que es reconociblemente una teoría física, implicar la existencia de la conciencia? Si esta teoría consiste en una descripción de la estructura y la dinámica de campos, ondas, partículas, etc., entonces todos los problemas usuales seguirán vigentes. No es claro que *cualquier* tipo de teoría física pueda ser suficientemente diferente de esto como para evitar las dificultades.

El problema es que los elementos básicos de las teorías físicas parecen reducirse siempre a dos cosas: la estructura y la dinámica de los procesos físicos. Diferentes teorías invocan diferentes tipos de estructuras. La física newtoniana invoca un espacio-tiempo euclidiano; la teoría de la relatividad invoca una variedad diferencial no euclidiana; la teoría cuántica invoca un espacio de Hilbert para funciones de onda. Y diferentes teorías invocan diferentes tipo de dinámicas dentro de esas estructuras: las leyes de Newton, los principios de la relatividad, las ecuaciones de onda de la mecánica cuántica. Pero de la estructura y la dinámica, sólo podemos obtener más estructura y dinámica. Esto hace posible satisfacer explicaciones de todo tipo de propiedades estructurales y funcionales de alto nivel, pero la experiencia consciente seguirá sin ser afectada. Ningún conjunto de hechos acerca de la estructura y dinámica de la física puede totalizar un hecho acerca de la fenomenología.

Por supuesto, existe un sentido en el cual la física del universo *debe* implicar la existencia de la conciencia, si *definimos* la física como la ciencia fundamental a partir de cuyos hechos y leyes surge todo lo demás. Esta interpretación de la física, sin embargo, trivializa la cuestión involucrada. Si permitimos que la física incluya a las teorías desarrolladas específicamente para tratar con el fenómeno de la conciencia, sin una motivación en consideraciones más básicas, entonces podremos obtener una "explicación" de la conciencia, pero ciertamente no será reductiva. Para nuestros propósitos, será mejor interpretar a la física como la ciencia fundamental desarrollada para explicar las observaciones del mundo externo. Si este tipo de física implicase los hechos sobre la conciencia, sin invocar a la propia conciencia en un papel crucial, entonces la estaría explicando reductivamente. Por los motivos que enuncié, sin embargo, hay

buenas razones para creer que ninguna explicación reductiva de este tipo es posible.

Casi todas las propuestas que existen relativas a la utilización de la física para explicar la conciencia se concentran en la parte más problemática de la física, la mecánica cuántica. Esto es comprensible: para que la física pueda explicar la conciencia se requeriría algo extraordinario, y la mecánica cuántica es de lejos, la parte más extraordinaria de la física contemporánea. Pero, a la larga, no parece ser lo suficientemente extraordinaria.

Por ejemplo, Penrose (1994) sugiere que la clave para la comprensión de la conciencia podría encontrarse en una teoría que reconcilia la teoría cuántica con la teoría general de la relatividad. Sugiere que efectos gravitacionales aun no comprendidos podrían ser responsables del colapso de la función de onda cuántica, lo que aportaría un elemento no algorítmico a las leyes de la naturaleza. Basándose en las ideas de Hameroff (1994), sugiere que la cognición humana podría depender de colapsos cuánticos en microtúbulos, que son estructuras proteínicas que se encuentran en la estructura de sostén de una neurona. Penrose y Hameroff sugieren que el colapso cuántico en los microtúbulos podría ser la base física de la experiencia consciente.

Estas ideas son extremadamente especulativas, pero es al menos *concebible* que puedan ayudar a explicar ciertos elementos del funcionamiento cognitivo humano. Penrose sugiere que el elemento no algorítmico en el colapso podría explicar ciertos aspectos de nuestra comprensión matemática, que él cree que van más allá de la capacidad de cualquier sistema algorítmico. Hameroff sugiere que el colapso de una función de onda superpuesta podría ayudar a explicar ciertos aspectos de la toma de decisiones humana. Pero nada de todo esto parece ayudar en la explicación de la experiencia consciente. ¿Por qué los procesos cuánticos en los microtúbulos deberían dar origen a la conciencia? La cuestión aquí es tan difícil como la cuestión correspondiente acerca de los procesos clásicos en un cerebro humano. Cuando se trata del problema de la experiencia, los procesos no algorítmicos y algorítmicos se encuentran en el mismo bote.

Algunos sugirieron que la *no localidad* de la mecánica cuántica, como lo sugieren experimentos recientes relacionados con la paradoja de Einstein-Podolsky-Rosen y el teorema de Bell, podría ser la clave de una teoría de la conciencia (véase Lahav y Shanks, 1992, para sugerencias según estas líneas). Pero, aunque la física sea no local, es difícil ver cómo esto podría ayudar en la explicación de la conciencia. Aun con un proceso físico no local, sigue siendo lógicamente posible

que el proceso pueda ocurrir en ausencia de la conciencia. La brecha explicativa es tan ancha como siempre.

La conexión más frecuentemente mencionada entre la conciencia y la mecánica cuántica se encuentra en el hecho de que en algunas interpretaciones de esta última, se necesita la medición de un observador consciente para producir el colapso de la función de onda. En este tipo de interpretación, la conciencia tiene un papel central en la dinámica de la teoría física. Estas interpretaciones son sumamente controversiales, pero de cualquier forma es notorio que no hacen nada para proporcionar una *explicación* de la conciencia. Más bien, simplemente *suponen* la existencia de la conciencia, y la utilizan para ayudar a explicar ciertos fenómenos físicos. Ocasionalmente se formulan teorías de la conciencia que hacen uso de esta relación (por ejemplo, Hodgson, 1988; Stapp, 1993), pero ciertamente no son teorías reductivas.[13]

No podemos descartar la posibilidad de que teorías físicas fundamentales como la mecánica cuántica desempeñen un papel clave en una teoría de la conciencia. Por ejemplo, quizá la conciencia resulte estar asociada a ciertas propiedades físicas fundamentales, o a ciertas configuraciones de aquellas propiedades, o quizás haya un vínculo aun más sutil. Pero, de cualquier forma, hay pocas esperanzas de que este tipo de teorías proporcione una *explicación* totalmente física de la conciencia. Cuando se trata de explicación reductiva, las teorías basadas en la física no están en una mejor situación que las teorías neurobiológicas o cognitivas.

## 6. La explicación evolutiva

Aun aquellos que toman a la conciencia seriamente suelen sentirse atraídos por la idea de una explicación evolutiva de la conciencia. Después de todo, la conciencia es una característica tan ubicua y fundamental que parece que debería haber surgido durante el proceso evolutivo por alguna *razón*. En particular, es natural suponer que surgió porque posee alguna función que no podría realizarse sin ella. Si pudiésemos tener una idea suficientemente clara de la función relevante, entonces tendríamos alguna idea de por qué la conciencia existe.

Desafortunadamente, esta idea sobrevalora lo que una explicación evolutiva podría proporcionarnos. El proceso de la selección natural no puede distinguir entre mi persona y la de mi gemelo zombi. La evolución selecciona propiedades según su papel funcional, y mi gemelo zombi realiza todas las funciones que yo realizo tan bien como

yo; en particular deja diseminadas tantas copias de sus genes como yo. Se deduce que la evolución sola no puede explicar por qué evolucionaron criaturas conscientes y no zombis.

Algunos podrían sentirse tentados a responder, "Pero un zombi *no puede* hacer todas las cosas que yo puedo hacer". Pero mi gemelo zombi es por definición físicamente idéntico a mí en su historia, así que ciertamente produce una conducta indistinguible. Cualquiera que desee cuestionar la capacidad de los zombis debe, entonces, encontrar algo erróneo en los argumentos al comienzo de este capítulo, en lugar de plantear sus objeciones aquí.

Para ver este punto de un modo diferente, nótese que el verdadero problema de la conciencia es explicar los principios en virtud de los cuales esta surge en los sistemas físicos. Es de suponer que estos principios —sean verdades conceptuales, necesidades metafísicas o leyes naturales— son constantes en el espacio y el tiempo: si una réplica física de mí hubiese surgido a la existencia hace un millón de años, habría sido exactamente tan consciente como yo. Los principios conectores mismos son por lo tanto independientes del proceso evolutivo. Aunque la evolución puede ser muy útil para explicar por qué evolucionaron sistemas físicos particulares, es irrelevante para la explicación de los principios puente en virtud de los cuales algunos de estos sistemas son conscientes.

## 7 ¿Hacia dónde va la explicación reductiva?

No es infrecuente que las personas estén de acuerdo con las críticas a las concepciones reductivas específicas, pero califiquen de algún modo esta aceptación: "Por supuesto *eso* no explica la conciencia, pero si esperamos un tiempo, alguna explicación surgirá". Espero que la presente discusión haya dejado en claro que los problemas de este tipo de explicación de la conciencia son más fundamentales que esto. Las dificultades con los modelos y teorías presentados aquí no se encuentran en los *detalles*; al menos, no tenemos necesidad de considerar los detalles para advertir lo que está mal en ellos. El problema se encuentra en la estrategia explicativa general. Estos modelos y teorías simplemente no son el *tipo* de cosas que puedan explicar la conciencia.

Es inevitable que se formulen "explicaciones" reductivas cada vez más sofisticadas de la conciencia, pero estas sólo producirán explicaciones crecientemente sofisticadas de las funciones cognitivas. Aun desarrollos "revolucionarios" como la apelación a redes conexionistas, dinámicas no lineales, vida artificial y mecánica cuántica sólo proporcionarán explicaciones funcionales más poderosas. Esto po-

dría encaminarse hacia una ciencia cognitiva muy interesante, pero el misterio de la conciencia no desaparecerá.

Cualquier concepción enunciada en términos puramente físicos padecerá el mismo problema. En última instancia, deberá enunciarse en términos de propiedades estructurales y dinámicas de procesos físicos, y no importa cuán sofisticada sea la concepción, sólo producirá más estructura y dinámica. Aunque esto es suficiente para tratar con la mayor parte de los fenómenos naturales, el problema de la conciencia va más allá de cualquier problema de estructura y función, de modo que se necesita un nuevo tipo de explicación.

Podría suponerse que eventualmente podría surgir una técnica explicativa reductiva que explicase cosas que no sean estructura y función, pero es muy difícil ver cómo esto sería posible, dado que las leyes de la física finalmente se formulan en términos de estructura y dinámica. La existencia de la conciencia será siempre un hecho ulterior en relación con los hechos estructurales y dinámicos, de modo que siempre permanecerá inexplicada por las concepciones físicas.

Debemos buscar en otro lugar, entonces, una explicación de la conciencia. No necesitamos abandonar el intento explicativo; sólo debemos abandonar la explicación *reductiva*. La posibilidad de explicar la conciencia no reductivamente permanece abierta. Este sería un tipo de explicación muy diferente, que requeriría de algunos cambios radicales en el modo como pensamos acerca de la estructura del mundo. Pero si realizamos estas modificaciones, los albores de una teoría de la conciencia podrán ser visibles a la distancia.

# 4
# El dualismo naturalista

## 1. Un argumento en contra del materialismo

En el capítulo anterior, me he ocupado de la pregunta explicativa "¿Puede la conciencia explicarse mediante teorías físicas?", en lugar de la pregunta ontológica, "¿Es física la conciencia?" Pero las dos están estrechamente relacionadas, y en este capítulo trazaré las consecuencias ontológicas de los argumentos del capítulo anterior. En particular, la no superveniencia lógica implica directamente que el materialismo es falso: existen otras características del mundo por encima y por debajo de las características físicas. El argumento básico de esto es como sigue.

1. En nuestro mundo, existen experiencias conscientes.
2. Hay un mundo lógicamente posible y físicamente idéntico al nuestro en el cual los hechos positivos acerca de la conciencia en nuestro mundo no son válidos.
3. Por lo tanto, los hechos acerca de la conciencia son hechos ulteriores acerca de nuestro mundo, por encima y por debajo de los hechos físicos.
4. El materialismo es falso.

Si un mundo zombi físicamente idéntico es lógicamente posible, se deduce entonces que la presencia de la conciencia es un hecho *extra* acerca de nuestro mundo, que no está garantizado por los hechos físicos solamente. La naturaleza de nuestro mundo no se agota en las características provistas por los hechos físicos; existen características extra debidas a la presencia de la conciencia. Para usar una frase debida a Lewis (1990), la conciencia contiene *información* fenoménica. Los hechos físicos constriñen de forma incompleta el modo de ser del mundo; los hechos acerca de la conciencia lo constriñen aún más.

Una conclusión similar puede obtenerse a partir de la posibilidad lógica de un mundo con experiencias conscientes *invertidas*. Un mundo de esta clase es físicamente idéntico al nuestro, pero algunos de los hechos acerca de la experiencia consciente en nuestro mundo no son válidos en ese mundo. Se deduce que los hechos acerca de la experiencia consciente en nuestro mundo son hechos ulteriores, más allá de los hechos físicos, y que el materialismo es falso.

De cualquiera de las dos formas, si la conciencia no es lógicamente superveniente a lo físico, entonces el materialismo es falso. La no superveniencia lógica implica que algún hecho positivo acerca de nuestro mundo no es válido en un mundo físicamente idéntico, de modo que es un hecho ulterior, más allá de los hechos físicos. Como en el capítulo 2, interpreto al materialismo como la doctrina de que los hechos físicos acerca del mundo agotan todos los hechos, en el sentido de que todo hecho positivo está implicado por los hechos físicos. Si los mundos zombi o los mundos invertidos son posibles, resultará entonces que los hechos físicos no implican a todos los hechos positivos acerca de nuestro mundo y, por lo tanto, el materialismo es falso.

Podemos usar aquí la imagen de Kripke. Cuando Dios creó el mundo, después de asegurarse de la validez de los hechos físicos, *todavía tuvo más trabajo que hacer*. Debía asegurarse de que los hechos acerca de la conciencia fueran válidos. La posibilidad de mundos zombi o mundos invertidos muestran que tenía una opción. El mundo podría haber carecido de experiencia, o podría haber contenido experiencias diferentes, aunque todos los hechos físicos hubiesen sido los mismos. Para asegurarse de que los hechos acerca de la conciencia son como son, hechos ulteriores debieron incluirse en el mundo.

## ¿Qué tipo de dualismo?

Este fracaso del materialismo lleva a una especie de *dualismo*: en el mundo hay características físicas y no físicas. La falsedad de la superveniencia lógica implica que la experiencia es fundamentalmente diferente en especie de cualquier característica física. Pero hay muchas variedades de dualismo, y es importante ver adónde exactamente nos lleva el argumento.

Los argumentos en el último capítulo establecen que la conciencia no es *lógicamente* superveniente a lo físico, pero esto no significa que no supervenga en absoluto. En los casos con los que estamos familiarizados, parece haber una dependencia sistemática de la experiencia consciente sobre la estructura física, y nada en los argumentos del último capítulo sugiere otra cosa. Sigue siendo tan plausible como siempre, por ejemplo, que si mi estructura física fuese

duplicada por alguna criatura en el mundo real, mi experiencia consciente también se duplicaría. De esta manera, sigue siendo plausible que la conciencia supervenga *naturalmente* a lo físico. Es este enfoque —superveniencia natural sin superveniencia lógica— el que desarrollaré.

Los argumentos no nos llevan a un dualismo como el de Descartes, con un reino separado de la sustancia mental que ejerce su propia influencia sobre los procesos físicos. La mejor evidencia de la ciencia contemporánea nos dice que el mundo físico está más o menos cerrado causalmente: para todo suceso físico, existe una causa física suficiente. Si esto es así, no hay lugar para un "fantasma en la máquina" mental que realice algún trabajo causal extra. Podría haber una pequeña excepción debido a la existencia de la indeterminación cuántica, pero argumentaré más adelante que es probable que esto no pueda utilizarse para asignarle un papel causal a una mente no física. En cualquier caso, debido a todos los argumentos del capítulo previo, sigue siendo plausible que los sucesos *físicos* puedan explicarse en términos físicos, de modo que un movimiento a un dualismo cartesiano sería una reacción más fuerte de lo necesario.

El dualismo implicado aquí es, en cambio, una especie de dualismo de *propiedades*: la experiencia consciente involucra propiedades de un individuo que no están implicadas por las propiedades físicas de ese individuo, aunque puede depender nomológicamente de esas propiedades. La conciencia es una *característica* del mundo más allá de sus características físicas. Esto no significa que sea una "sustancia" aparte; la cuestión de qué se requeriría para justificar un dualismo de sustancias me parece bastante poco clara. Todo lo que sabemos es que en este mundo hay propiedades de individuos —las propiedades fenoménicas— que son ontológicamente independientes de las propiedades físicas.

Existe un tipo más débil de dualismo de propiedades con el que este enfoque no debería ser confundido. A veces se dice que el dualismo de propiedades se aplica a cualquier dominio cuyas propiedades no sean propiedades invocadas por la física o directamente reducibles a tales propiedades. En este sentido, la aptitud biológica no es una propiedad física. Pero este tipo de "dualismo" es una variedad demasiado débil. No hay nada que ontológicamente sea *fundamentalmente* nuevo en propiedades como la aptitud, ya que estas son todavía lógicamente supervenientes a las propiedades microfísicas. El dualismo de propiedades de esta variedad es por entero compatible con el materialismo. En contraste, el dualismo de propiedades que propongo involucra características fundamentalmente nuevas del mundo. Debido a que estas propiedades ni siquiera son lógicamente

supervenientes a las propiedades microfísicas, son no físicas en un sentido mucho más fuerte. Cuando hablo de dualismo de propiedades y de propiedades no físicas, es este enfoque más fuerte y el sentido más fuerte de no fisicalismo lo que tengo en mente.

Sigue siendo plausible, sin embargo, que la conciencia *surja* de una base física, aun cuando no esté *implicada* por esa base. La posición en la que nos encontramos es que la conciencia surge a partir de un sustrato físico en virtud de ciertas leyes contingentes de la naturaleza que no están ellas mismas implicadas por leyes físicas. Muchas personas que se consideran a sí mismas materialistas sostienen implícitamente esta posición. Es común escuchar, "Por supuesto soy materialista; la mente seguramente surge del cerebro". La propia presencia de la palabra "surge" debería resultarnos una advertencia. No tendemos a decir "el aprendizaje surge del cerebro", por ejemplo, y si lo hiciésemos, sería en un sentido temporal de "surge". En su lugar, diríamos más naturalmente que el aprendizaje *es* un proceso en el cerebro. El hecho de que la mente necesita *surgir* del cerebro indica que ocurre algo ulterior, más allá de los hechos físicos.[1]

Algunas personas pensarán que el enfoque debería ser una versión de materialismo más que de dualismo, porque plantea una dependencia legaliforme fuerte de los hechos fenoménicos respecto de los hechos físicos, y porque el dominio físico sigue siendo autónomo. Por supuesto tiene poco sentido discutir acerca de un nombre, pero me parece que la existencia de hechos contingentes ulteriores, además de los hechos físicos, es una modificación suficientemente significativa de la perspectiva recibida del mundo materialista como para merecer un rótulo diferente. Por cierto, si todo lo que se requiriese en el materialismo fuese que todo hecho esté conectado de un modo legaliforme con hechos físicos, entonces el materialismo se volvería una doctrina débil.

Aunque es una variedad de dualismo, no hay nada anticientí- fico o sobrenatural en este enfoque. El mejor modo de pensar en él es como sigue. La física postula un número de características *fundamentales* del mundo: espacio-tiempo, masa-energía, carga, spin, etc. También plantea un número de leyes fundamentales en virtud de las cuales estas características fundamentales están relacionadas. Estas últimas no pueden explicarse en términos de características más básicas, y las leyes fundamentales no pueden explicarse en términos de leyes más básicas; simplemente deben ser tomadas como primitivas. Una vez que las leyes fundamentales y la distribución de las características fundamentales han sido definidas, sin embargo, casi todo en el mundo surge de ellas. Es por ello que una teoría fundamental en la física se conoce a veces como una "teoría de todo". Pero el hecho de que la

conciencia no superviene a las características físicas nos muestra que esta teoría física no es totalmente una teoría de todo. Para llevar a la conciencia dentro del alcance de una teoría fundamental, necesitamos introducir *nuevas* propiedades y leyes fundamentales.

En su libro *Dreams of a Final Theory* (1992), el físico Steven Weinberg hace notar que lo que hace que una teoría fundamental de la física sea especial es que lleva a una cadena explicativa desde el comienzo al fin que a la postre explica todo. Pero se ve forzado a conceder que una teoría de este tipo podría no explicar la conciencia. Cuanto más, dice, podemos explicar los "correlatos objetivos" de la conciencia. "Eso no será una explicación de la conciencia, pero estará muy cerca" (p. 45). Pero, por supuesto, no lo suficientemente cerca. No explica todo lo que ocurre en el mundo. Para ser consistentes, debemos reconocer que una teoría verdaderamente final necesita un componente adicional.

Hay dos maneras en las que esto podría hacerse. Podríamos tomar a la propia experiencia como una característica fundamental del mundo, junto con el espacio y el tiempo, el spin, la carga, etc. Esto es, ciertas propiedades fenoménicas deberán ser consideradas propiedades *básicas*. Alternativamente, quizás exista alguna *otra* clase de propiedades fundamentales nuevas de las que las propiedades fenoménicas se derivan. Los argumentos previos mostraron que estas no pueden ser propiedades físicas, pero es posible que sean propiedades no físicas de una nueva variedad, respecto de las cuales las propiedades fenoménicas son lógicamente supervenientes. Estas propiedades estarían relacionadas con la experiencia, del mismo modo como las propiedades físicas básicas están relacionadas con propiedades no básicas como la temperatura. Podríamos llamar a estas propiedades *protofenoménicas*, ya que no son en sí mismas fenoménicas, pero juntas pueden producir lo fenoménico. Por supuesto es muy difícil imaginar qué podría ser una propiedad protofenoménica, pero no podemos descartar la posibilidad de que exista. La mayor parte del tiempo, sin embargo, hablaremos como si las propiedades fundamentales fueran ellas mismas fenoménicas.

Allí donde tenemos nuevas propiedades fundamentales, también tenemos nuevas leyes fundamentales. Aquí las leyes fundamentales serán leyes *psicofísicas*, que especifican cómo las propiedades fenoménicas (o protofenoménicas) dependen de propiedades físicas. Estas leyes no interferirán con las leyes físicas; estas últimas ya forman un sistema cerrado. En cambio, serán *leyes de superveniencia*, que nos cuentan cómo surge la experiencia a partir de los procesos físicos. Hemos visto que la dependencia de la experiencia sobre lo

físico no puede derivarse de las leyes físicas, de modo que cualquier teoría final debe incluir leyes de esta variedad.

Por supuesto, en esta etapa, tenemos muy poca idea acerca de qué apariencia tendrá la teoría fundamental relevante, o cuáles serán las leyes psicofísicas fundamentales. Pero tenemos razones para creer que una teoría de este tipo existe. Hay buenas razones para creer que existe una relación legaliforme entre los procesos físicos y la experiencia consciente, y cualquier relación legaliforme debe apoyarse en las leyes fundamentales. El estudio de la física nos cuenta que las leyes fundamentales son típicamente simples y elegantes; deberíamos esperar lo mismo de las leyes fundamentales de una teoría de la conciencia. Una vez que tenemos una teoría fundamental de la conciencia que acompañe a una teoría fundamental de la física, podríamos verdaderamente tener una teoría de todo. Dadas las leyes físicas y psicofísicas básicas, y dada la distribución de las propiedades fundamentales, podemos esperar que todos los hechos acerca del mundo seguirán. Desarrollar una teoría de esta clase no será sencillo, pero debería ser posible en principio.

En cierto sentido, lo que ocurre aquí con la conciencia es análogo a lo que ocurrió con el electromagnetismo en el siglo XIX. Se realizaron intentos de explicar los fenómenos electromagnéticos en términos de las leyes físicas ya comprendidas, que involucraban principios mecánicos y otros similares, pero esto no fue exitoso. Resultó que para explicar los fenómenos electromagnéticos debían considerarse como fundamentales características como la carga y las fuerzas electromagnéticas; Maxwell introdujo nuevas leyes electromagnéticas fundamentales. Sólo de este modo pudieron ser explicados los fenómenos. De la misma forma, para explicar la conciencia no son suficientes las características y leyes de la teoría física. Para una teoría de la conciencia se necesitan nuevas características y leyes fundamentales.

Este enfoque es totalmente compatible con una cosmovisión científica contemporánea, y además es totalmente naturalista. Según este enfoque, el mundo todavía consiste de una red de propiedades fundamentales relacionadas por leyes básicas, y todo debe ser en última instancia explicado en esos términos. Lo que ocurrió es que el inventario de propiedades y leyes ha sido expandido, como sucedió con Maxwell. Más allá de esto, nada en este enfoque contradice ninguna cosa en la teoría física; más bien, complementa dicha teoría. Una teoría física es una teoría de procesos físicos, y una teoría psicofísica nos dice como esos procesos dan origen a la experiencia.

Para captar el espíritu del enfoque que propongo, lo llamo *dualismo naturalista*. Es naturalista porque plantea que todo es una

consecuencia de una red de propiedades y leyes básicas, y porque es compatible con todos los resultados de la ciencia contemporánea. Y como con las teorías naturalistas en otros dominios, este enfoque nos permite *explicar* la conciencia en términos de las leyes naturales básicas. No es necesario que haya nada especialmente trascendente acerca de la conciencia; sólo es otro fenómeno natural. Lo único que ocurrió es que nuestra imagen de la naturaleza se expandió. A veces se interpreta el "naturalismo" como sinónimo de "materialismo", pero me parece que la aceptación de una comprensión naturalista del mundo puede sobrevivir al fracaso del materialismo. (Si algún lector lo duda, señalo al resto de este trabajo como evidencia.) Algunos podrían encontrar una cierta ironía en el nombre de este enfoque, pero lo más importante es que transmite el mensaje fundamental: adoptar el dualismo no es necesariamente adoptar el misterio.

En cierto modo, aquellos que sostienen este tipo de dualismo podrían estar temperamentalmente más cerca de los materialistas que de los dualistas de otras variedades. Esto se debe, en parte, a que evita cualquier elemento trascendental y a su aceptación de la explicación natural y, en parte, a su aceptación de la causalidad física de la conducta. Recíprocamente, al evitar cualquier compromiso con un fantasma en la máquina, este enfoque evita las peores inverosimilitudes de los enfoques dualistas tradicionales. Frecuentemente escuchamos que los éxitos de la ciencia cognitiva y la neurociencia hacen que el dualismo no sea plausible, pero no todas las variedades de dualismo están afectadas por igual. Todos estos éxitos se basan en explicaciones físicas de la conducta y de otros fenómenos físicos, de modo que no distinguen entre el enfoque materialista y el dualista naturalista.

Dos notas finales. Algunos se preguntarán por qué, si la experiencia es fundamental, no puede ser una propiedad *física*. Después de todo, ¿no es la física justamente la ciencia de lo que es verdaderamente fundamental? La respuesta: Seguramente, si *definimos* a la física de ese modo, la experiencia será una propiedad física, y las leyes de superveniencia serán leyes de la física. Pero en una interpretación más natural de la "física" y lo "físico", la experiencia no califica como integrante. La experiencia no es una propiedad fundamental que los físicos necesiten formular en su teoría del mundo externo; la física forma una teoría cerrada y consistente aun sin la experiencia. Dada la posibilidad de un mundo zombi, existe un sentido claro en el cual la experiencia es superflua para la física tal como se la entiende usualmente. Es más natural, por lo tanto, considerar la experiencia como una propiedad fundamental que no es una propiedad física, y considerar las leyes psicofísicas como leyes fundamentales de la

naturaleza que no son leyes de la física. Pero nada importante gira en torno de la cuestión terminológica, en tanto se mantenga la claridad de la forma del enfoque.

También debería notarse que aunque llamo a mi enfoque una variedad de dualismo, es posible que pueda resultar ser una especie de monismo. Es posible que lo físico y lo fenoménico resulten ser dos aspectos diferentes de una sola clase abarcativa, de un modo similar a como la materia y la energía resultan ser dos aspectos de un mismo tipo. Nada de lo que he dicho permite desechar esta idea y, de hecho, tengo una cierta simpatía por ella. Pero sigue vigente que si una variedad de monismo es verdadera, no puede ser un monismo *materialista*, debe ser algo más amplio.

## Objeciones

Podrían plantearse algunas objeciones al argumento en contra del materialismo que formulé al comienzo de este capítulo. Algunas de estas son objeciones a la premisa (2), la negación de la superveniencia lógica; traté las objeciones de este tipo en el capítulo previo. Aquí trataremos las objeciones al paso de la no superveniencia lógica a la falsedad del materialismo. Las objeciones más serias de este tipo son las que invocan una necesidad *a posteriori*. Las trataré en la próxima sección. Aquí consideraremos algunas objeciones menores.

A veces se argumenta que la conciencia podría ser una propiedad *emergente*, en un sentido que es todavía compatible con el materialismo. En trabajos recientes sobre sistemas complejos y la vida artificial, suele sostenerse que las propiedades emergentes no son predecibles a partir de las propiedades de bajo nivel, pero que de todas formas son físicas. Ejemplos de esto son el surgimiento de la autoorganización en los sistemas biológicos, o el surgimiento de patrones de bandada a partir de reglas simples en pájaros simulados (Langton, 1990; Reynolds, 1987). Pero las propiedades emergentes de esta clase no son análogas a la conciencia. Lo que es interesante en estos casos es que las propiedades relevantes no son consecuencias obvias de las leyes de bajo nivel; pero todavía son lógicamente supervenientes a los *hechos* de bajo nivel. Si se dan *todos* los hechos físicos acerca de un sistema de esta clase a lo largo del tiempo, entonces la circunstancia de que ocurra la autoorganización será derivable de ellos de forma directa. Esto es exactamente lo que esperaríamos, ya que propiedades como la autoorganización y la agrupación en bandadas son funcionales o estructurales.

Si la conciencia es una propiedad emergente, lo es en un sentido mucho más fuerte. Existe una noción más fuerte de emergencia,

utilizada por los emergentistas británicos (por ejemplo, Broad, 1925), según la cual las propiedades emergentes no son ni siquiera predecibles a partir del conjunto completo de hechos físicos de bajo nivel. Es razonable decir (como los emergentistas británicos lo hacían) que la experiencia consciente es emergente en este sentido. Pero parece mejor considerar este tipo de emergencia como una variedad del dualismo de propiedades. A diferencia de los ejemplos más "inocentes" de emergencia que dimos antes, la variedad fuerte requiere de nuevas leyes fundamentales para que puedan surgir las propiedades emergentes.

Otra objeción es que la conciencia y lo físico podrían ser dos aspectos de lo mismo, en el sentido en que la estrella matutina y la estrella vespertina son dos aspectos de Venus. Si esto es así, la conciencia podría ser física en cierto sentido. Pero, nuevamente, debemos preguntar: ¿El aspecto fenoménico está implicado por el aspecto físico? Si lo está, tenemos una variedad de materialismo, pero volvemos a los argumentos del capítulo 3. Si no lo está, entonces el aspecto fenoménico proporciona contingencias ulteriores en el mundo más allá del aspecto físico, y la dualidad de los aspectos nos determina una especie de dualismo de propiedades. Podría ser posible que la dualidad de lo físico y lo fenoménico puedan subsumirse bajo un monismo más amplio, pero este no será un monismo de lo físico exclusivamente.

Una tercera objeción es sugerida por el trabajo de Searle (1992). Como yo, Searle sostiene que la conciencia es sólo naturalmente superveniente a lo no físico. Acepta que una réplica zombi es lógicamente posible, y sostiene que la conciencia es meramente *causada* por los estados del cerebro. Pero rechaza que esto sea una variedad de dualismo, ni siquiera dualismo de propiedades. Esto podría parecer una mera cuestión terminológica, pero Searle insiste que la condición ontológica de la conciencia es la misma que la de características físicas como la liquidez, de modo que la cuestión no es *meramente* terminológica. El argumento de Searle de que el enfoque no es dualista consiste en que algo similar ocurre en cualquier otro aspecto: por ejemplo, el $H_2O$ causa liquidez, pero nadie es un dualista acerca de la liquidez.

Sin embargo, me parece evidente que esta es una falsa analogía. Dados todos los hechos microfísicos acerca de un lote particular de $H_2O$, es lógicamente imposible que esos hechos puedan ser válidos sin que la liquidez haya sido instanciada. La noción de una réplica no líquida de un lote de $H_2O$ líquido es simplemente incoherente. Se deduce que la relación entre los hechos microfísicos y la liquidez es mucho más estrecha que una simple relación causal. Las caracterís-

ticas microfísicas no *causan* liquidez; la *constituyen*. Esto es enteramente diferente de lo que ocurre en el caso de la conciencia, de modo que la analogía no funciona. La conciencia es ontológicamente nueva de un modo mucho más significativo que la liquidez.[2]

Finalmente, algunos hallarán el argumento que formulé en defensa del dualismo reminiscente del dado por Descartes. Este filósofo afirmaba que podía imaginar su mente existiendo separadamente de su cuerpo, de modo que su mente no podía ser idéntica a su cuerpo. Por lo general se considera que este tipo de razonamiento es erróneo: tan sólo de que podamos imaginarnos que A y B no sean idénticos, no se deduce que A y B no lo sean (piense en la estrella matutina y la estrella vespertina, por ejemplo). ¿No podría mi argumento cometer un error similar? El mundo zombi sólo muestra que es *concebible* que podamos tener un estado físico sin conciencia; no muestra que un estado físico y la conciencia no son idénticos.

Esto es comprender erróneamente el argumento, sin embargo. Es crucial que, tal como la formulé, la argumentación no dependa de cuestiones de *identidad* sino de *superveniencia*. La forma del argumento no es "Podemos imaginarnos un estado físico *P* sin conciencia, por lo tanto la conciencia no es un estado físico *P*" sino, más bien, "Podemos imaginarnos que *todos* los hechos físicos son válidos sin que lo sean los hechos acerca de la conciencia, de modo que los hechos físicos no agotan todos los hechos". Esta es una especie totalmente diferente de argumento. En general, las argumentaciones modales en favor del dualismo formuladas en términos de identidad son menos concluyentes que las formuladas en términos de superveniencia; esta es una razón de por qué realicé una planteo enteramente en términos de superveniencia, y evité casi por completo hablar de identidad. Me parece que las cuestiones sobre la superveniencia son las fundamentales aquí.

No obstante podríamos intentar responder a este argumento con una estrategia análoga a la respuesta en contra de Descartes. Por ejemplo, podríamos notar que mi estrategia todavía se basa en una especie de inferencia de la conceptibilidad a la posibilidad, lo que podría ser cuestionable. Consideraré las estrategias de este tipo en el apartado siguiente.

## 2. Objeciones a partir de la necesidad *a posteriori**

Una respuesta popular a este tipo de argumentación es objetar que sólo demuestra que el mundo zombi es *lógicamente* posible, lo que es bastante diferente de que sea *metafísicamente* posible. Mientras que la coherencia conceptual es suficiente para la posibilidad

lógica, la posibilidad metafísica está más restringida. También suele formularse la cuestión sugiriendo que hay una diferencia entre la *conceptibilidad* y la verdadera *posibilidad*. Aunque puede ocurrir que un mundo zombi sea concebible, se requiere algo más para mostrar que es posible en el sentido metafísico relevante para la falsedad del materialismo.

Esta objeción suele estar acompañada de una apelación a *Naming and Necessity*, de Kripke (1980), donde se demuestra la existencia de verdades necesarias como "El agua es $H_2O$", cuya necesidad es sólo cognoscible *a posteriori*. En los términos de estos objetantes, es lógicamente posible que el agua no sea $H_2O$, pero no es metafísicamente posible. De un modo similar, no sería demasiado extraño suponer que los zombis puedan ser lógicamente posibles pero metafísicamente imposibles. De ser así, quizás esto sea suficiente para salvar al materialismo.

Esta es, de lejos, la estrategia más común de los materialistas que están convencidos de que no existe ninguna implicación entre los conceptos físicos y los fenoménicos. Según este enfoque, puede existir una brecha conceptual sin una brecha metafísica. El enfoque ofrece la atractiva perspectiva de tomar en serio a la conciencia a la vez que conserva el materialismo. Desafortunadamente, en un examen detenido del mismo pueden verse de modo bastante directo sus falencias. La noción de una necesidad *a posteriori* no soporta el peso que este argumento requiere, y de hecho es una especie de falso camino en este contexto.[3]

Esto puede verse si utilizamos el marco bidimensional para tratar con la necesidad *a posteriori* que desarrollamos en el capítulo 2, apartado 4. Recuérdese que en este marco existen dos intensiones (funciones de mundos posibles en referentes) asociadas con cualquier concepto: una intensión primaria (determinada *a priori*) que fija la referencia en el mundo real, y una intensión secundaria (determinada *a posteriori*) que selecciona la referencia en mundos contrafácticos. La intensión primaria asociada con "agua" es algo así como "sustancia acuosa". La intensión secundaria es "$H_2O$", que se deriva de la intensión primaria aplicando el operador *dthat* de Kaplan: "*dthat* (sustancia acuosa)" selecciona al $H_2O$ en todos los mundos posibles, así como la sustancia acuosa es $H_2O$ en el mundo real.

La "posibilidad lógica" se reduce a la verdad posible de un enunciado cuando se evalúa de acuerdo con las intensiones primarias involucradas (lo que llamé posibilidad-1 en el capítulo 2). Las intensiones primarias de "agua" y "$H_2O$" difieren, de modo que es lógicamente posible en este sentido que el agua no sea $H_2O$. La "posibilidad metafísica" se reduce a la verdad posible de un enunciado

cuando se evalúa de acuerdo con las intensiones secundarias involucradas (esto es, posibilidad-2). Las intensiones secundarias de "agua" y "$H_2O$" son una misma, de modo que es metafísicamente necesario que el agua sea $H_2O$.

La objeción, por lo tanto, se reduce a la cuestión de que mediante argumentos a partir de la conceptibilidad y otros similares, lo que demostramos es la posibilidad de un mundo zombi utilizando las intensiones *primarias* de las nociones involucradas, pero no las más apropiadas intensiones *secundarias.* Aunque la intensión primaria de las nociones fenoménicas puede no corresponder a la de ninguna noción física, las intensiones secundarias pueden ser las mismas. Si esto es así, entonces los conceptos fenoménicos y físicos/funcionales pueden seleccionar las mismas propiedades *a posteriori* a pesar de la distinción *a priori.* Una objeción de este tipo podría hacerla un preconizador del "psicofuncionalismo" (véase Block, 1980), que equipara las propiedades fenoménicas a las propiedades funcionales *a posteriori,* o un defensor de un enfoque que equipara las propiedades fenoménicas a ciertas propiedades neurofisiológicas *a posteriori.*

El modo más fácil de ver que nada de esto afecta el argumento en favor del dualismo es notar que mi argumentación sigue siendo válida aunque en todo momento nos concentremos en la intensión primaria e ignoremos la intensión secundaria. Vimos en el capítulo 2 que la intensión primaria es la más relevante para la explicación, pero también es útil para la argumentación en favor del dualismo. Porque, ya sea que coincidan o no la intensión primaria y secundaria, la primaria determina una propiedad perfectamente apropiada de objetos en mundos posibles. La propiedad de ser sustancia acuosa es una propiedad perfectamente razonable, aun cuando no es lo mismo que la propiedad de ser $H_2O$. Si podemos mostrar que hay mundos posibles que son físicamente idénticos al nuestro pero en los cuales la propiedad introducida por la intensión primaria está ausente, entonces surgirá el dualismo.

Esto es justamente lo que hicimos con la conciencia. Hemos visto que hay mundos exactamente como el nuestro que carecen de conciencia, de acuerdo con su intensión primaria. Esta diferencia en los mundos es suficiente para mostrar que estas son propiedades de nuestro mundo más allá de las propiedades físicas. Por analogía, si pudiésemos mostrar que hay mundos físicamente idénticos al nuestro en los cual no existe una sustancia acuosa, habríamos establecido un dualismo respecto del agua tan bien como si hubiésemos establecido que son mundos físicamente idénticos al nuestro en los cuales no hay $H_2O$. Y, lo que es importante, la diferencia con respecto a la intensión primaria puede establecerse independientemente de facto-

res *a posteriori*, de modo que las consideraciones acerca de la necesidad *a posteriori* son irrelevantes.

(Dos notas técnicas aquí: Estrictamente hablando, una intensión primaria determina una propiedad *relativa a un centro* de un objeto en un mundo posible (o una relación entre objetos y centros), ya que la intensión primaria se aplica a mundos posibles *centrados*. Pero esta relatividad no puede utilizarse para ayudar a nuestro objetante. Una vez que se especifica la localización de un centro, una intensión primaria determina una propiedad no indicativa perfectamente apropiada; y todos los argumentos del capítulo 3 son válidos, aun cuando la localización del centro esté incluida en la base de superveniencia. Por ejemplo, aunque los hechos de María acerca del mundo incluyen hechos acerca de dónde está localizada, esto no la habilitará para conocer cómo es ver el color rojo.

Podríamos también preocuparnos por el hecho de que el concepto de conciencia no está presente en el centro del mundo zombi, mientras que la aplicación de una intensión primaria podría requerir la presencia del concepto relevante en el centro del mundo. (¡Incluso podríamos comenzar a preocuparnos acerca de la aplicación del concepto *zombi*!) Creo que la situación es más sutil que esto —las intensiones primarias no necesitan requerir la presencia del concepto original— pero, en cualquier caso, podemos fácilmente eludir esta preocupación si consideramos un mundo *parcialmente* zombi: uno en el que yo estoy en el centro, consciente, con todos los conceptos relevantes, pero en el cual algunas otras personas son zombis.)

La irrelevancia de la necesidad *a posteriori* puede apoyarse además en la observación de que en el caso de la conciencia, la intensión primaria y secundaria coinciden. Lo que se requiere para que un estado sea una experiencia consciente en el mundo real es que posea una sensación fenoménica, y lo que se requiere para que algo sea una experiencia consciente en un mundo contrafáctico es que posea una sensación fenoménica. La diferencia entre la intención primaria y secundaria del concepto agua refleja el hecho de que podría haber algo que parezca y se sienta como agua en algún mundo contrafáctico que de hecho no sea agua, sino meramente sustancia acuosa. Pero si algo se siente como una experiencia consciente, aun en algún mundo contrafáctico, *es* una experiencia consciente. Lo único que significa ser una experiencia consciente, en cualquier mundo posible, es tener una cierta sensación (Kripke señala un punto similar, aunque él lo formula en términos de propiedades esenciales más que en términos de significado).

Aun si alguien insistiese en que las intensiones primaria y secundaria difieren, sin embargo, el argumento seguiría siendo

válido. Simplemente nos concentramos en la intensión primaria utilizada para fijar la referencia, como más arriba. Por ejemplo, si la "conciencia" se reduce a "*dthat* (tiene una sensación fenoménica)", entonces simplemente nos concentramos en la intensión "tiene una sensación fenoménica". Los argumentos en el capítulo 3 establecen que existe un mundo posible en el cual mi réplica carece de una sensación fenoménica, de modo que la propiedad de tener una sensación fenoménica es un hecho más allá de los hechos físicos, y la argumentación en favor del dualismo es exitosa.[4]

El modo más general de establecer la cuestión es notar que nada en la necesidad *a posteriori* de Kripke hace que algún mundo lógicamente posible resulte imposible. Simplemente nos dice que algunos de ellos son descriptos erróneamente, porque estamos aplicando términos de acuerdo con sus intensiones primarias en lugar de con las más apropiadas intensiones secundarias. Podríamos haber pensado que era posible *a priori* que el agua fuese XYZ, en lugar de $H_2O$. Al concebir esto, imaginamos algo así como un mundo en el cual XYZ es el líquido que se encuentra en océanos y lagos. Sin embargo, el análisis de Kripke nos muestra que debido al modo como resulta el mundo actual, estamos describiendo erróneamente este mundo como uno en el que XYZ es agua, debido a que lo estamos describiendo con la intensión primaria en lugar de la más apropiada intensión secundaria. Estrictamente hablando, es un mundo en el cual XYZ es sustancia acuosa. Estas consideraciones no pueden demostrar la imposibilidad de este mundo aparentemente posible; tan sólo muestran el modo correcto de describirlo.

Como vimos en el capítulo 2, las consideraciones de Kripke nos indican que la intensión secundaria $F_a$: W → R a veces difiere de la intensión primaria $f$: W* → R. Esto plantea algunas restricciones *a posteriori* sobre las condiciones de aplicación de los conceptos, pero el espacio relevante de mundos permanece constante en todo momento; la única diferencia entre los argumentos de las dos funciones involucra la localización de un centro. De modo que, aunque puede haber dos clases de posibilidades de *enunciados*, sólo hay una clase relevante de posibilidades de *mundos*.

Se deduce que si existe un mundo concebible que es físicamente idéntico al nuestro pero carece de ciertas características positivas de nuestro mundo, entonces ninguna consideración acerca de la designación de términos como "conciencia" puede hacer nada para descartar la posibilidad metafísica del mundo. Podemos simplemente olvidarnos de la semántica de esos términos y notar que el mundo posible relevante claramente carece de *algo*, lo llamemos o no "conciencia". En el mejor de los casos las consideraciones kripkeanas podrían

decirnos cómo debería describirse apropiadamente ese mundo y sus características relevantes, pero no tienen ningún efecto sobre su posibilidad; y la mera posibilidad de un mundo semejante, independientemente de cómo se describa, es todo lo que el argumento en favor del dualismo necesita para tener éxito.

### Una estrategia alternativa

Existe un modo bastante diferente en el que podríamos recurrir a la necesidad *a posteriori* para evitar el dualismo. Podría argumentarse que sostener que el mundo zombi es *físicamente idéntico* al nuestro es describirlo erróneamente. Del mismo modo que el mundo XYZ parece contener agua pero no es así, el mundo zombi *parece* ser físicamente idéntico pero es físicamente diferente. Esto puede sonar extraño, pero hay un modo de traducirlo. Un opositor podría argumentar que hay propiedades esenciales para la constitución física del mundo que no son accesibles a la investigación física. Al concebir un mundo "físicamente idéntico", en realidad sólo estamos concibiendo un mundo que es idéntico desde el punto de vista de la investigación física, pero que difiere en las propiedades esenciales inaccesibles, que son también las propiedades que garantizan la conciencia.

Por ejemplo, podría ocurrir que para que algo califique como electrón en un mundo contrafáctico, no es suficiente que esté causalmente relacionado con otras entidades físicas al modo como un electrón lo está. Podría también requerirse alguna esencia oculta de los electrones. En ese enfoque, el concepto de un electrón es algo así como "*dthat* (la entidad que desempeña el papel de electrón)". La referencia a los electrones se fija mediante una caracterización extrínseca, pero luego es rigidificada de modo que las entidades con la misma naturaleza intrínseca son seleccionadas en mundos contrafácticos, independientemente de que desempeñen el papel apropiado, y de esta forma las entidades que desempeñan el papel en esos mundos no califican como electrones a menos que tengan la naturaleza intrínseca apropiada. Lo mismo sucedería para propiedades como masa, que podría entenderse como "*dthat* (la propiedad que desempeña el papel de masa)". La naturaleza esencial de los electrones o de la masa estaría entonces oculta para la teoría física, que caracteriza a los electrones y a la masa sólo de modo extrínseco. Si esto es así, podría ocurrir que las propiedades esenciales relevantes sean propiedades fenoménicas o protofenoménicas, de modo que su instanciación podría garantizar la existencia de la conciencia en nuestro mundo.

Si este fuese el caso, el mundo zombi que estamos concibiendo carecería de estas propiedades esenciales ocultas y por lo tanto no

sería físicamente idéntico a nuestro mundo. El mundo zombi daría los mismos resultados que nuestro mundo cuando se lo evalúa de acuerdo con las intensiones primarias de los predicados físicos, que se aplican sobre la base de relaciones extrínsecas, pero no cuando se evalúan de acuerdo con las intensiones secundarias, que requieren la esencia oculta. Dado esto, la experiencia consciente podría supervenir "metafísicamente" a las propiedades físicas después de todo. (Un argumento muy similar a este es ofrecido por Maxwell [1978] y también sugerido por el enfoque de Lockwood [1989]. En la formulación de Maxwell, la idea básica es que aun cuando los conceptos fenoménicos no pueden recibir análisis neutrales de tópico que seleccionan propiedades físicas subyacentes, los conceptos físicos sí pueden recibir análisis neutrales de tópico que podrían seleccionar propiedades fenoménicas subyacentes.)[5]

Esta es, en muchas formas, una objeción más interesante que la anterior. Ciertamente se basa en una metafísica especulativa, pero esto no le impide ser una posición coherente. Una respuesta más directa es que se basa en un enfoque incorrecto de la semántica de los términos físicos. Los predicados físicos se aplican incluso *a posteriori* sobre la base de relaciones extrínsecas entre entidades físicas, independientemente de cualquier propiedad oculta. Esta es una cuestión puramente conceptual: si los electrones en nuestro mundo tuviesen propiedades protofenoménicas ocultas, ¿llamaríamos electrón a una entidad contrafáctica que es en todos los respectos idéntica salvo que carece de esas propiedades? Creo que sí. No sólo la referencia a los electrones se fija por el papel que estos desempeñan en una teoría; el propio concepto de electrón se define mediante ese papel, lo que determina la aplicación del concepto a través de mundos. La noción de un electrón que tiene todas las propiedades extrínsecas de los protones actuales no parece ser coherente, y tampoco la noción de que existe un mundo en el cual la masa desempeña el papel que actualmente tiene la carga. La concepción semántica dada más arriba predice que estas nociones deberían ser coherentes; es, entonces, una concepción falsa de los conceptos.

Las intuiciones semánticas pueden diferir, pero, como es usual, hay una respuesta que es más profunda que las intuiciones semánticas. Aun si admitiésemos que ciertas propiedades ocultas pudiesen ser constitutivas de las propiedades físicas, la diferencia entre este enfoque y el dualismo de propiedades que defiendo es pequeña. Sigue siendo válido que el mundo tiene propiedades fenoménicas que no están determinadas por las propiedades que la física revela. Luego de asegurarse de que un mundo es idéntico al nuestro desde el punto de vista de nuestras teorías físicas, Dios debe realizar todavía un

esfuerzo más para hacer que ese mundo sea idéntico al nuestro en todo aspecto. El dualismo de propiedades "físicas" y "no físicas" se reemplaza en este enfoque por un dualismo de propiedades físicas "accesibles" y "ocultas", pero la cuestión esencial sigue siendo la misma.

Volveré más adelante al punto de vista de que las entidades físicas tienen una naturaleza protofenoménica intrínseca, pero la metafísica de la perspectiva sigue siendo la misma independientemente del enfoque que adoptemos hacia la semántica de los predicados físicos. Como antes, las intensiones secundarias y la necesidad *a posteriori* sólo marcan una diferencia semántica, no metafísica. Como sea que se especifique el enfoque, este admite propiedades fenoménicas o protofenoménicas como fundamentales, por lo que está más cerca de una versión de dualismo (o quizás un idealismo o un monismo neutral, como veremos más adelante) que de una versión de materialismo.

### Necesidad metafísica fuerte

El análisis bidimensional que acabamos de analizar establece que una invocación a la necesidad *a posteriori* kripkeana no tiene ninguna fuerza en contra del argumento a partir de la superveniencia. Este tipo de necesidad no plantea restricciones *a posteriori* en el espacio de mundos posibles, sino que simplemente restringe el modo como se utilizan ciertos términos para describirlos; de manera que si hay un mundo lógicamente posible que es idéntico al nuestro en todos los aspectos físicos pero no en todos los aspectos positivos, entonces estas consideraciones no pueden contradecir la posibilidad metafísica del mundo.

Algunos podrían sostener, sin embargo, que los mundos relevantes podrían ser, *no obstante*, metafísicamente imposibles. Sería posible sostener que existe una modalidad de posibilidad metafísica que es distinta y más restringida que la posibilidad lógica, y que surge por razones independientes de las consideraciones kripkeanas. De acuerdo con este enfoque, existen menos mundos metafísicamente posibles que mundos lógicamente posibles, y la necesidad *a posteriori* de ciertos enunciados puede surgir de factores relativamente independientes de la semántica de los términos involucrados. Podemos llamar a esta modalidad hipotética *necesidad metafísica fuerte*, en oposición a la *necesidad metafísica débil* introducida por el marco kripkeano.

Según esta perspectiva, existen mundos que son completamente concebibles, incluso de acuerdo con las restricciones más fuertes sobre la conceptibilidad, pero que no son posibles en absoluto. Esta es

una brecha entre la conceptibilidad y la posibilidad mucho más fuerte que cualquier otra brecha que pueda hallarse en cualquier otro lugar. Hay un sentido en el cual la verdad de *enunciados* como “El agua es XYZ” es concebible pero no posible, pero estos ejemplos nunca desechan la posibilidad de ningún *mundo* concebible. Son meramente instancias en las que un mundo de esta clase se describe de un modo erróneo. La necesidad metafísica fuerte va más allá de esto. Según esta posición, “mundo zombi” puede describir correctamente el mundo que estamos concibiendo, incluso según una intensión secundaria. El único problema es que el mundo no es metafísicamente posible.[6]

La respuesta breve a esta objeción es que no hay ninguna razón para creer que exista una modalidad semejante. Estas “necesidades metafísicas” plantearán restricciones sobre el espacio de mundos posibles que son primitivas e inexplicables. Podría ser razonable permitir hechos primitivos e inexplicables en *nuestro* mundo, pero la existencia de estos hechos en el espacio de mundos posibles sería bastante extravagante. El dominio de lo posible (en oposición al domino de lo natural) no tiene lugar para este tipo de restricción arbitraria.

La posición no puede sostenerse por analogía, ya que no hay ninguna analogía disponible.[7] Hemos visto que las analogías con la necesidad de “El agua es $H_2O$”, “Héspero es Fósforo”, etc., fracasan, ya que esos ejemplos requieren un único espacio de mundos. Si algunos mundos son lógicamente posibles pero metafísicamente imposibles, parecería que nunca podríamos saberlo. Por definición la información no está disponible *a priori* y la información *a posteriori* sólo nos dice acerca de *nuestro* mundo. Esto puede servir para localizar nuestro mundo en el espacio de mundos posibles, pero es difícil ver cómo podría darnos información acerca de la extensión de ese espacio. Cualquier aseveración acerca de restricciones agregadas de posibilidad metafísica parecería ser una cuestión de estipulación arbitraria; de la misma manera podríamos estipular que es metafísicamente imposible que una piedra pueda moverse hacia arriba cuando la soltamos.

Además, la posición lleva a una proliferación *ad hoc* de modalidades. Si la aceptásemos, deberíamos admitir *cuatro* tipos de enunciados de posibilidad y necesidad, aun dejando de lado la modalidad natural: posibilidad y necesidad de acuerdo con las intensiones primaria o secundaria sobre el espacio de mundos lógica o metafísicamente posibles. Y si consideramos la posibilidad de mundos en lugar de enunciados, ahora tendríamos *tres* clases objetivas de mundos posibles: mundos lógicamente posibles, mundos metafísica-

mente posibles y mundos naturalmente posibles. Tenemos buenas razones para creer en la primera y la última de estas clases, pero tenemos pocas razones para creer en una tercera clase distinta como metafísicamente dada.

Alguien que sostenga que un mundo zombi es lógicamente posible pero metafísicamente imposible tiene que responder la pregunta clave: *¿Por qué Dios no podría haber creado un mundo zombi?* Supuestamente, estaba dentro de los poderes de Dios, cuando creó el mundo, hacer cualquier cosa que fuera lógicamente posible. Sin embargo, el defensor de la necesidad metafísica debe decir que la posibilidad es coherente, pero que Dios no podría haberla creado, o que Dios podría haberla creado, pero de todas formas es metafísicamente imposible. Lo primero es bastante injustificado, y lo segundo es totalmente arbitrario. De cualquier forma, si lo segundo fuese válido, el argumento en contra del materialismo todavía estaría vigente; después de fijar los hechos físicos acerca del mundo, Dios todavía tenía trabajo por hacer.

Aun si se aceptase este enfoque, se parecería mucho al dualismo de propiedades que defiendo en varios aspectos cruciales. Según esta perspectiva, la existencia de la conciencia todavía no puede derivarse del conocimiento físico, de modo que la conciencia no puede explicarse reductivamente. Entonces seguiríamos necesitando ciertos principios conectores primitivos para explicar la superveniencia de lo fenoménico a lo físico. La única diferencia entre los enfoques es que los principios psicofísicos relevantes son considerados "leyes de necesidad" primitivas en lugar de leyes de la naturaleza. Para todo propósito *explicativo* en la construcción de una teoría, nos encontramos en la misma posición en la que nos deja el dualismo de propiedades; la diferencia principal está en una estipulación ontológica.

La única motivación real de esta perspectiva parecería ser salvar al materialismo a toda costa, quizá debido a los problemas percibidos con el dualismo. Pero este tipo de materialismo parece mucho más misterioso que la alternativa dualista. La invocación de principios primitivos "metafísicamente necesarios" que constriñen el espacio de mundos posibles introduce un elemento mucho más problemático y ciertamente mucho menos naturalista, que la mera invocación de leyes naturales ulteriores postuladas por la propiedad de dualismo. Al fin de cuentas, la invocación de un nuevo grado de necesidad es un tipo de solución por estipulación *ad hoc* que plantea tantos problemas como respuestas. El enfoque salva al materialismo sólo al costo de hacer que resulte totalmente misterioso cómo la conciencia *podría* ser física.[8]

## Las limitaciones cognitivas

Existe una última posición que podría adoptar un materialista que encuentre que el mundo zombi es concebible pero quiera todavía salvar al materialismo. En la posición que analizamos más arriba, el materialista acepta que la noción de un zombi es totalmente coherente, aun para un ser máximamente racional, pero de todas formas niega su posibilidad metafísica, lo que lleva así a un cuadro "de dos capas" de los mundos lógica y metafísicamente posibles. Pero un materialista podría también argumentar que la aparente conceptibilidad surge de algún tipo de racionalidad deteriorada, de modo que si sólo fuésemos más inteligentes veríamos que la descripción del mundo no es, después de todo, coherente. En este enfoque, el mundo no es ni siquiera lógicamente posible; sólo ocurre que las limitaciones de las facultades cognitivas humanas nos llevan erróneamente a creer que lo es. (Esta podría ser una interpretación de la posición de McGinn, [1991].)

Podríamos intentar darle apoyo a esta posición por analogía con la necesidad de ciertas verdades matemáticas complejas que van más allá de nuestros poderes de comprensión matemática. Si nuestros poderes matemáticos son computables, dichas verdades deben existir (por el teorema de Gödel), e incluso si no, bien podrían existir de todos modos. (Quizá la conjetura de Goldbach sea un ejemplo, o quizá la hipótesis del continuo o su negación.) Estas verdades son necesarias aun cuando no sean cognoscibles *a priori* por nosotros, ni tampoco estén fundamentadas en una combinación de factores cognoscibles *a priori* y empíricos al modo de las necesidades kripkeanas. ¿La implicación de los hechos físicos en los hechos fenoménicos podría ser una necesidad de este tipo, de algún modo más allá de nuestros poderes de comprensión modal?[9]

Sin embargo, la analogía es imperfecta. En el caso matemático, nuestro razonamiento modal deja la cuestión *abierta*; nuestras intuiciones de conceptibilidad no nos dicen nada en un sentido o en el otro. Puede haber algún sentido débil en el cual sea "concebible" que los enunciados sean falsos —por ejemplo, son falsos según todo lo que sabemos— pero este no es un sentido que produzca un mundo concebible cuando fallan. En el caso zombi, en cambio, la cuestión no queda abierta: parece haber un mundo claramente concebible en el cual la implicación es falsa. Para salvar el materialismo, la posibilidad de este mundo debe ser desechada a pesar de la mejor evidencia de nuestros poderes modales; pero nada en el caso matemático se acerca a proporcionar un ejemplo en el cual un mundo aparentemente posible es desechado de ese modo. Una vez más, cualquier brecha

entre la conceptibilidad y la posibilidad que el materialista pudiese invocar aquí debe ser *sui generis*, no sustentada por analogías relevantes en ningún otro lado.[10]

Por supuesto, un materialista podría aceptar esto y defender un limitación cognitiva *sui generis*. Para hacerlo, debería sostener que los argumentos en el capítulo 3 están todos equivocados en modos que no podemos apreciar. Además de requerir que la racionalidad imperfecta haga que nuestras intuiciones de conceptibilidad se descarríen masivamente, el enfoque también requiere que una versión más inteligente de María pudiese saber cómo es ver el color rojo sobre la base de información física, y que exista un análisis de conceptos fenoménicos que sustente la implicación de hechos físicos en hechos fenoménicos (quizás un análisis estructural o funcional), aunque su exactitud se encontraría más allá de nuestros poderes de apreciación.

Aunque debe concederse que cualquier argumento filosófico *podría* estar equivocado debido a una limitación cognitiva, en ausencia de alguna razón sustancial para creer en ello, este tipo de objeción parece bastante *ad hoc*. Como antes, la motivación principal parecería ser el deseo de aferrarse al materialismo a cualquier costo. Una opción de este tipo debería ser siempre la *última* opción que consideremos, sólo después de haber abandonado los *argumentos* sustanciales que señalan adónde nos hemos equivocado y los intentos sustanciales por desarrollar una alternativa al materialismo. Si encontramos una alternativa sustancial que sea satisfactoria, entonces cualquier motivación para este punto de vista desaparecerá.[11]

## 3. Otros argumentos en favor del dualismo*

No soy el primero en utilizar en contra del materialismo el argumento a partir de la posibilidad lógica.[12] Ciertamente, creo que de una forma u otra es el argumento antimaterialista fundamental en la filosofía de la mente. Sin embargo, no ha recibido la cuidadosa atención que merece. Una mayor atención se concentró en los dos argumentos antimaterialistas de Jackson (1982) y Kripke (1972). Estos argumentos me parece que están relacionados, pero, según mi opinión, son menos fundamentales. El de Jackson es importante por el acceso que proporciona al argumento a partir de la superveniencia lógica, y la porción más convincente del de Kripke depende, como veremos, de este mismo argumento.

Analizamos anteriormente el argumento de Jackson, basado en el conocimiento, en el contexto de establecer la no superveniencia lógica, donde tenía un papel de apoyo. Recuérdese que el argumento se ocupa de María, una neurocientífica criada en una habitación en blanco y negro, que conoce todos los hechos físicos acerca del procesamiento del color en el cerebro. Más tarde, cuando por primera vez ve un objeto rojo, aprende algunos nuevos hechos. En particular, aprende cómo es ver el color rojo. El argumento concluye que los hechos físicos no agotan todos los hechos y por lo tanto que el materialismo es falso.

Este argumento está estrechamente relacionado con los argumentos de los zombis y del espectro invertido, en el sentido que todos giran en torno de que los hechos fenoménicos no están implicados por los hechos físicos. En cierto modo, son dos caras de un mismo argumento. Como argumentación directa en contra del materialismo, sin embargo, suele considerarse que el argumento de Jackson es vulnerable debido a su utilización de la noción intensional del conocimiento. Muchos ataques al argumento se concentraron en su intensionalidad, sosteniendo, por ejemplo, que el mismo hecho puede ser conocido de dos modos diferentes. Estos ataques fracasan, creo, pero el modo más directo de verlo es proceder directamente a la no superveniencia, que se formula en términos de metafísica más que de epistemología. El marco teórico que desarrollé ayuda a sacar a la luz exactamente por qué las diversas objeciones no son exitosas. Analizaré algunas de estas objeciones en lo que sigue.

Primero, varios críticos argumentaron que aunque María obtiene nuevo conocimiento al ver el color rojo, este conocimiento no corresponde a ningún *hecho* nuevo. Simplemente, llega a conocer un hecho viejo de una forma nueva, bajo un nuevo "modo de presentación", debido a la intensionalidad del conocimiento (Churchland, 1985; Horgan, 1984b; Lycan, 1995; McMullen, 1985; Papineau, 1993; Teller, 1992; Tye, 1986). Por ejemplo, Tye y Lycan recurrieron a la diferencia intensional entre "Este líquido es agua" y "Este líquido es $H_2O$": en un sentido los dos enunciados expresan el mismo hecho, pero se puede conocer uno sin conocer el otro. De un modo similar, Churchland recurre a la brecha entre el conocimiento de la temperatura y el conocimiento de la energía cinética media, Horgan discute la diferencia entre el conocimiento sobre Clark Kent y el conocimiento sobre Superman, mientras que McMullen señala a Mark Twain y Samuel Clemens.

Estas brechas surgen precisamente debido a la diferencia entre las intensiones primaria y secundaria. Podemos conocer cosas acerca

del agua sin saber cosas acerca del $H_2O$ porque las intensiones primarias difieren: no hay ninguna conexión *a priori* entre los pensamientos sobre el agua y los pensamientos sobre el $H_2O$. Sin embargo, en un sentido, sólo hay un conjunto de hechos acerca de los dos: debido a la identidad *a posteriori* entre el agua y el $H_2O$, las intensiones secundarias relevantes coinciden. (No es obvio que *debamos* individualizar los hechos de esta manera, de forma que los hechos sobre el agua y sobre el $H_2O$ sean los mismos hechos, pero lo aceptaré a propósito del hilo de la argumentación.)[13] En la terminología utilizada antes, “Si esto es agua, es $H_2O$” es lógicamente contingente pero metafísicamente necesario. Esta objeción, por lo tanto, se reduce precisamente a lo mismo que la objeción a partir de la distinción entre la necesidad lógica y la necesidad metafísica (kripkeana) discutida antes, y el análisis que realizamos en ese momento sobre las intensiones primarias y secundarias es suficiente para refutarla.

Podemos también formular la cuestión de un modo más directo. Siempre que conozcamos un hecho bajo un modo de presentación pero no bajo otro, existirá un hecho *diferente* del que no tenemos conocimiento, un hecho que conecta los dos modos de presentación.[14] Si sabemos que Héspero es visible pero no que Fósforo es visible (porque no sabemos que Héspero es Fósforo), entonces no sabemos que un mismo objeto es la estrella más brillante en el cielo matutino y la estrella más brillante en el cielo vespertino. Este es un hecho separado del que no tenemos conocimiento en absoluto. De modo similar, si sabemos que Superman puede volar pero no que Clark Kent puede volar, entonces no sabemos que hay un individuo que es periodista de *El Planeta* y que usa una capa. Si sabemos que el agua es húmeda pero no que el $H_2O$ lo es, no sabemos que la sustancia en los lagos está hecha de moléculas de $H_2O$. Y así siguiendo.

Más formalmente. Digamos que “$a$ es G” y “$b$ es G” son el mismo hecho en este sentido, pero no podemos conectar los dos hechos *a priori*. Esto debe ocurrir porque $a = b$ y las intensiones secundarias son las mismas, pero las intensiones primarias son diferentes: quizás $a$ sea equivalente a $dthat(P)$ y $b$ a $dthat(Q)$. Si sabemos que $a$ es $G$ pero no que $b$ es $G$, entonces carecemos del conocimiento fáctico de que algo es $P$ y $Q$. Más en general, carecemos del conocimiento fáctico de que algo es $P'$ y $Q'$, donde estas son descripciones identificadoras cualesquiera tal que sabemos que $a$ es $P'$ y $b$ es $Q'$. Este hecho es bastante diferente de los hechos que poseíamos inicialmente. Aun cuando se lo interprete de acuerdo con las intensiones secundarias, habrá un mundo posible en el cual $a$ es $F$ pero en el cual nada es $P$ y $Q$ al mismo tiempo (o $P'$ y $Q'$).

(Como en el apartado 2, existe la complicación de que *P* y *Q* pueden ser propiedades relativas a índices, pero esto no cambia nada fundamental. Para hacer que el hecho nuevo desconocido sea estrictamente no indicador, sólo necesitamos cambiarlo por el hecho "Existe un punto [con propiedad *X*] en el cual *P* y *Q* seleccionan la misma cosa". *X* es un respaldo para el caso de que sepamos que en *otros* lugares *P* y *Q* seleccionan la misma cosa; en esta circunstancia, simplemente hacemos que *X* sea lo suficientemente específico como para distinguirlo de esas otras localizaciones. El caso extremo en el que carecemos de todo autoconocimiento distintivo se reduce al caso indicativo puro, que analizaremos más abajo.)

Se deduce que si María adquiere algún conocimiento fáctico del que anteriormente carecía —aunque sólo sea el conocimiento de un hecho viejo bajo un modo diferente de presentación— entonces debe haber algún hecho verdaderamente novedoso del que obtiene conocimiento. En particular, debe llegar a saber un nuevo hecho que *involucra* ese modo de presentación. Dado que ella ya sabía todos los hechos físicos, se deduce que el materialismo es falso. Los hechos físicos no son exhaustivos en ningún sentido.

Esta respuesta puede parecer menos directa que la correspondiente al argumento a partir de la posibilidad lógica. El marco de superveniencia elimina la cuestión menos precisamente definida de cómo individualizar fragmentos de conocimiento y hace, entonces, que la discusión sea menos confusa. Aún así, un análisis detenido muestra que las analogías agua-$H_2O$ y las objeciones relacionadas fracasan de todas formas. A pesar del hecho de que esta es fácilmente la respuesta más popular al argumento del conocimiento, es también la más débil de las respuestas principales. Simplemente no resiste el análisis.

Una segunda objeción más sofisticada, debida a Loar [1990], también sostiene que María adquiere nuevo conocimiento de hechos viejos debido a la intensionalidad, pero explícitamente va más allá de las analogías usuales agua-$H_2O$. Loar reconoce que las analogías con los ejemplos usuales no son útiles para el materialista, ya que (en nuestra terminología) esas analogías permiten que las nociones físicas y fenoménicas tengan intensiones primarias distintas, y el antimaterialista puede simplemente aplicar el argumento a la propiedad que corresponde a la intensión primaria. Como dice Loar, aun cuando "calor" y algún predicado de la mecánica estadística *designen* la misma propiedad (intensión secundaria), de todas formas *introducen* propiedades distintas (intensión primaria). Por esta razón lleva el argumento aun más lejos, y sostiene que los dos predicados pueden introducir la misma propiedad —esto es, compartir la misma

intensión primaria— aun cuando esta igualdad no sea cognoscible *a priori*. Si esto es así, entonces el conocimiento que María posee de las propiedades fenoménicas puede ser sólo conocimiento de propiedades físicas/funcionales, aun cuando ella puede no haberlas conectado de antemano.

Pero, ¿cómo pueden dos intensiones primarias coincidir sin que seamos capaces de saberlo *a priori*? Sólo si el espacio de mundos posibles es más pequeño de lo que hubiésemos pensado *a priori*. Creemos que las intensiones difieren porque concebimos un mundo en el que tienen diferente referencia, tal como un mundo zombi. La posición de Loar, por lo tanto, requiere que este mundo no sea realmente posible, a pesar del hecho de que no podemos descartarlo sobre una base conceptual, y a pesar del hecho de que la necesidad *a posteriori* kripkeana no nos es de utilidad. Esta posición, por lo tanto, se reduce precisamente a lo mismo que la objeción de la "necesidad metafísica fuerte" que consideramos más arriba. Como esa objeción, la posición de Loar requiere que un condicional de los hechos físicos a los hechos fenoménicos sea metafísicamente necesario a pesar de ser lógicamente contingente, pero esta brecha no puede explicarse por una diferencia en las intensiones primarias. Como esa objeción, la posición de Loar requiere una restricción primitiva y arbitraria sobre los mundos posibles. Loar no ofrece ningún argumento en favor de esa restricción, y su posición está sujeta precisamente a las mismas críticas.[15]

Podríamos esperar que haya una tercera objeción análoga a la "estrategia alternativa" intermedia del apartado 2. Esta se traduciría en la aseveración de que María no conoce verdaderamente todos los hechos físicos. Conoce todos los hechos formulados en los términos de la física, pero carece del conocimiento acerca de las esencias ocultas (fenoménicas o protofenoménicas) de las entidades físicas. Si ella tuviese este conocimiento, entonces conocería los hechos fenoménicos. Como antes, sin embargo, este enfoque sólo puede tener una pretensión muy débil al nombre de "materialismo". Como mi propia perspectiva, este enfoque debe tomar a las propiedades fenoménicas o protofenoménicas como propiedades fundamentales.

Una cuarta objeción traza una conexión entre la dificultad de María y una ausencia de conocimiento indicativo (Bigelow y Pargetter, 1990; Mc Mullen, 1985; Papineau, 1993). Aunque María adquiere nuevo conocimiento, se argumenta que esto no es más problemático que otros casos en los que alguien que conoce todos los hechos objetivos relevantes descubre algo nuevo: por ejemplo, un amnésico omnisciente que descubre "*Yo* soy Rudolf Lingens", o un bien informado insomne que no sabe que son las 3,49 de la mañana *ahora* (véase

Perry, 1979, y Lewis, 1979). En estos casos, existe una brecha entre el conocimiento físico y el conocimiento indicativo, así como hay una brecha entre el conocimiento físico y el conocimiento fenoménico en el caso de María.

La conexión podría realizarse de dos modos. Primero, un objetante podría tratar de *reducir* el conocimiento fenoménico a conocimiento indicativo, argumentando que lo único que María no posee es conocimiento de esta clase. Segundo, podría tratar de trazar una *analogía* entre los dos casos, argumentando que en el caso indicativo la brecha epistémica no lleva a una brecha ontológica (la indicatividad no falsifica el materialismo), de modo que el caso fenoménico no tiene por qué llevarnos tampoco a una brecha ontológica.

La estrategia de reducción fracasa claramente. Como vimos en el capítulo 2, las expresiones indicativas, al igual que los hechos acerca de la experiencia consciente, no son lógicamente supervenientes a los hechos físicos, pero se determinan mediante el agregado de un "hecho indicativo" débil acerca de la localización del agente en cuestión. Pero, aun cuando le otorguemos a María un conocimiento perfecto acerca de su relación indicativa con el resto del mundo físico, su conocimiento de las experiencias de rojo no mejorará en lo más mínimo. Al no poseer conocimiento fenoménico, su carencia es mucho mayor que la de alguien que no posee conocimiento indicativo.

La estrategia de la analogía es más interesante. Podríamos responder argumentando en favor de una brecha ontológica incluso en el caso indicativo (véase, por ejemplo, Nagel, 1983), pero existe una respuesta más directa. Para verlo, nótese que en el caso indicativo, un argumento análogo al del apartado 1 no alcanza a despegar: no existe ningún mundo no centrado concebible en el cual los hechos físicos sean los mismos que en el nuestro, pero en el cual los hechos indicativos difieran. En los mundos no centrados, los hechos indicativos ni siquiera se aplican. Con seguridad, existe un mundo *centrado* concebible relevante, pero son los mundos no centrados los que son relevantes a la cuestión ontológica. (Si no, existiría una brecha ontológica también en el caso indicativo, de modo que el argumento del objetante no podría iniciarse.)[16] Así, sólo en este caso podemos eliminar explicativamente la brecha epistémica notando que las conexiones epistémicas están determinadas por intensiones primarias *centradas*, mientras que la conexiones ontológicas están determinadas por propiedades que corresponden a intensiones no centradas. Esto se refleja en la única grieta que encontramos en el argumento del apartado 2, y en el argumento análogo en este apartado: el hecho de que las intensiones primarias determinan sólo propiedades relativas a un centro. Esta abertura sólo deja pasar un único

ítem de conocimiento indicativo irreducible (la localización del centro de un mundo centrado) sin ningún costo ontológico, pero nada más. Una vez que la localización de un centro está especificada, la grieta se cierra. Los hechos fenoménicos permanecen indeterminados aun cuando la localización de un centro esté especificada, de manera que la experiencia consciente queda abandonada a su suerte.[17]

Si un materialista desea aferrarse al materialismo, debe sostener que María no hace ningún *descubrimiento* acerca del mundo en absoluto. El materialismo requiere la superveniencia lógica, y esta exige que María no pueda obtener ningún conocimiento fáctico nuevo de ninguna clase cuando experimenta el color rojo por primera vez. De esta manera, en una quinta estrategia, Lewis (1990) y Nemirow (1990) argumentan que María adquiere, cuanto más, una nueva *destreza*. Por ejemplo, la destreza de imaginar la visión de cosas rojas y reconocerlas cuando las ve. Pero, esto es sólo conocimiento de *cómo*, no conocimiento de *qué*. Cuando ella experimenta el color rojo por primera vez, no aprende ningún hecho acerca del mundo.[18]

A diferencia de las opciones previas, esta estrategia no sufre de problemas internos. Su dificultad principal es que resulta sumamente inverosímil. Sin duda, María adquiere algunas destrezas cuando experimenta el color rojo por primera vez, así como adquiere algunas destrezas cuando aprende a andar en bicicleta. Pero, en el primer caso, parece aprender algo más: algunos *hechos* sobre la naturaleza de la experiencia. En base a todo lo que sabía con anterioridad, la experiencia de las cosas rojas podría haber sido como esto o podría haber sido como aquello, o podría no haber sido como ninguna otra cosa. Pero ahora sabe que es como *esto*. Ella limitó el espacio de posibilidades epistémicas. Ningún conocimiento nuevo de esta clase ocurre cuando un mecánico omnisciente aprende a andar en bicicleta (excepto, quizá, por el conocimiento acerca de la fenomenología de andar en bicicleta). De modo que esta respuesta fracasa en dar cuenta de lo que ocurre cuando María aprende cómo es ver el color rojo.

Podemos también usar métodos más indirectos para ver que el descubrimiento de María involucra al conocimiento fáctico. Por ejemplo, Loar (1990) señala que este tipo de conocimiento puede estar inmerso en condicionales: "Si ver cosas rojas es así y ver cosas azules es de otro modo, entonces ver cosas púrpuras probablemente sea de esta manera"; "Si es así para los perros ver el color rojo, entonces se deduce tal y tal cosa", etc. Otro ejemplo; como Lycan (1995) señala,[19] lo que imaginamos puede resultar correcto o erróneo; entonces, después de ver unos pocos colores, María podría imaginar cómo es ver otro color, y lo que imagina podría ser correcto o incorrecto. Si esto es

así, entonces conocer cómo es algo es conocer una verdad acerca del mundo y el análisis en base a la destreza fracasa.

Dennett (1991) adopta una posición relacionada pero más extrema, argumentando que María no aprende nada en absoluto. Hace notar que María podría usar su conocimiento neurofisiológico para reconocer, al ver un objeto rojo, que es de ese color notando sus efectos sobre sus reacciones, que pueden diferir de los efectos producidos por algo azul. (Si un equipo de experimentadores intenta engañarla mostrándole una manzana azul, ella podría no dejarse engañar.) Quizás esto sea así, pero todo lo que se deduce es que *contra* Lewis y Nemirow, María tenía ciertas destrezas de reconocimiento aun antes de haber tenido su primera experiencia del color rojo. Esto no muestra en absoluto que María tenía el conocimiento crucial: conocimiento de cómo sería ver el color rojo. Eso sólo ocurriría si ya hubiésemos aceptado el análisis de destreza de "saber cómo es"; pero si hubiésemos aceptado ese análisis, el argumento en contra del materialismo ya habría sido derrotado. De modo que el argumento de Dennett es, aquí, un camino equivocado.

En última instancia, la estrategia que un materialista *debe* adoptar es negar que María adquiera conocimiento acerca del mundo. Y el único modo defendible de hacerlo parece ser por medio de un análisis de destreza de "saber cómo es". Esta es la única posición con la coherencia interna como para no ser rebatida por objeciones técnicas, así como el funcionalismo analítico es finalmente el modo más coherente en que un materialista puede resistirse al argumento a partir de la superveniencia lógica. Pero, en contraposición, la propia inverosimilitud de la negación de que María adquiera conocimiento acerca del mundo es evidencia de que el materialismo está condenado.[20]

Hemos visto que el argumento modal (el argumento a partir de la posibilidad lógica) y el argumento del conocimiento son dos caras de la misma moneda. Creo que, en principio, cada uno tiene éxito independientemente, pero en la práctica funcionan mejor en tándem.[21] Si tomamos sólo el argumento del conocimiento: la mayoría de los materialistas encuentran difícil negar que María adquiera conocimiento acerca del mundo, pero suelen negar el paso desde allí al fracaso del materialismo. Si tomamos sólo el argumento modal: la mayoría de los materialistas encuentra difícil negarlo a partir de la conceptiblidad de los zombis o del espectro invertido al fracaso del materialismo, pero suelen negar la premisa. Si se reúnen ambas argumentaciones, el argumento modal fortalecerá el argumento del conocimiento donde este necesita ayuda, y viceversa. En la combina-

ción posiblemente más poderosa de las dos argumentaciones podemos usar el argumento del conocimiento para establecer de modo convincente la no superveniencia lógica y el argumento modal para dar el paso desde la no superveniencia a la falsedad del materialismo de una forma persuasiva.

## El argumento de Kripke

El argumento de Kripke estaba dirigido a la tesis de identidad formulada por Place (1957) y Smart (1959), pero puede verse que tiene una gran fuerza en contra de todas las formas de materialismo. Analizaré las fortalezas y debilidades de este argumento con cierto detalle y llegaré a la conclusión de que las partes que tienen éxito son precisamente aquellas partes que corresponden al argumento de la superveniencia lógica.

La argumentación es aproximadamente la siguiente. De acuerdo con la tesis de identidad, ciertos estados mentales (como los dolores) y estados cerebrales (como la descarga de las fibras C) son idénticos, aun cuando "dolor" y "descarga de fibras C" no *significan* la misma cosa. Originalmente se suponía que la identidad era contingente, no necesaria, del mismo modo que la identidad entre el agua y el $H_2O$ es contingente. En contra de esto, Kripke argumenta que todas las identidades son necesarias: siempre que los términos *X* e *Y* tienen una designación rígida y seleccionan el mismo individuo o clase a través de los mundos, si *X* es *Y*, entonces *X* es *necesariamente Y*. El agua es *necesariamente* $H_2O$, argumenta; esto es, el agua es $H_2O$ en todo mundo posible. La identidad puede *parecer* contingente —esto es, podría parecer que existe un mundo posible en el cual el agua no es $H_2O$ sino XYZ— pero esto es ilusorio. De hecho, el mundo posible que estamos imaginando no contiene agua en absoluto. Es sólo un mundo en el que existe alguna sustancia acuosa —sustancia que se parece y se comporta como agua— hecha de XYZ. Al afirmar que esta sustancia es agua, la estamos describiendo erróneamente.

Del mismo modo, argumenta Kripke, si los dolores son idénticos a la descarga de las fibras C, entonces esta identidad debe ser necesaria. Pero la identidad no *parece* ser necesaria. A primera vista, podemos imaginar un mundo posible en el que los dolores ocurren sin ningún estado cerebral (dolor incorpóreo), y podemos imaginar un mundo en el que las fibras C producen descargas sin ningún dolor acompañante (en un zombi, digamos). Además, argumenta, estas posibilidades no pueden eliminarse explicando que sólo son posibilidades aparentes, al modo como se eliminó la posibilidad de agua sin $H_2O$. Para que ese fuese el caso, tendríamos que estar *describien-*

*do erróneamente* el mundo del "dolor incorpóreo" como uno en el cual ocurre el dolor, cuando en realidad sólo existe "sustancia dolorosa" (algo que se siente como dolor). De un modo similar, deberíamos estar describiendo erróneamente a un zombi como alguien que no tiene dolor, cuando lo único que en realidad no tiene es sustancia dolorosa. En una concepción de este tipo, el zombi supuestamente tendría dolor real, que es la descarga de las fibras C; pero este no se experimentaría como verdadero dolor.

Sin embargo, según Kripke, este no puede ser el caso: *todo lo que se necesita* para que algo sea dolor es que se experimente como dolor. No existe ninguna distinción entre dolor y sustancia dolorosa, al modo como sí existe una distinción entre agua y sustancia acuosa. Podríamos tener algo que se experimente como agua sin que sea agua, pero no podemos tener algo que se experimente como dolor sin que sea dolor. La sensación del dolor es *esencial*. De modo que la posibilidad de los dolores sin el estado cerebral (y viceversa) no puede descartarse como antes. Esos mundos posibles realmente son posibles, y los estados mentales no son necesariamente idénticos a los estados cerebrales. Se deduce que no pueden ser en absoluto idénticos a estados cerebrales.

Kripke hace funcionar el argumento de dos modos diferentes, una vez en contra de las teorías de identidad de ejemplares y otra en contra de las teorías de identidad de tipos. Las teorías de identidad de ejemplares sostienen que los dolores *particulares* (como mi dolor ahora) son idénticos a estados cerebrales particulares (como el de las fibras C que descargan en mi cabeza en este momento). Kripke argumenta en la forma de más arriba que un dolor particular podría ocurrir sin un estado cerebral particular asociado y viceversa, de modo que no pueden ser idénticos. Las teorías de identidad de tipos sostienen que los estados mentales y los estados cerebrales son idénticos como *tipos*: el dolor, por ejemplo, podría ser idéntico como tipo a la descarga de las fibras C. Kripke sostiene que esto puede refutarse directamente por el hecho de que podemos instanciar el tipo estado mental sin el tipo estado cerebral y viceversa. En síntesis, podemos contar aquí cuatro argumentos separados, divididos según el objetivo (teorías de identidad de ejemplares o de tipos) y según el método del argumento (desde la posibilidad de incorporeidad o desde la posibilidad de los zombis).

Existen algunas diferencias obvias entre el argumento de Kripke y el argumento que formulé. Primero, el argumento de Kripke está planteado enteramente en términos de identidad, mientras que yo utilicé la noción de superveniencia. Segundo, el argumento de Kripke está estrechamente ligado a su aparato teórico que involucra

designadores rígidos y necesidad *a posteriori*, mientras que dicho aparato sólo tiene un papel secundario en mi argumentación, para responder ciertas objeciones específicas. Tercero, suele considerarse que el argumento de Kripke se apoya en un cierto esencialismo acerca de diversos estados, mientras que yo no invoco ninguna doctrina similar en mi argumento. Cuarto, mi argumento no recurre en ningún lugar a la posibilidad de incorporeidad, como lo hace Kripke. Sin embargo, existen obvias similitudes. Ambos son argumentos modales, que asignan papeles claves a la necesidad y a la posibilidad. Y ambos recurren a la posibilidad lógica de disociar estados físicos de los estados fenoménicos asociados.

Analizaré ahora qué es lo que tiene éxito y qué es lo que falla en los argumentos de Kripke, partiendo de aquellos en contra de la identidad de ejemplares. Suele sostenerse que estos son dudosos. Esto se debe principalmente a que se basan en intuiciones acerca de lo que puede considerarse *esa misma cosa* a través de mundos posibles, ya que esas intuiciones son notablemente no fiables. La aseveración de Kripke de que podríamos tener *ese mismo* estado de dolor sin *ese mismo* estado cerebral se basa en la aseveración de que lo que es esencial en el estado de dolor es su sensación y sólo su sensación. Pero, afirmaciones de este tipo acerca de las propiedades esenciales de individuos son difíciles de justificar. El teórico de la identidad ejemplar puede responder argumentando que es igualmente plausible que la descarga de las fibras C sea una propiedad esencial del estado. Por supuesto, la descarga de fibras C no parece ser esencial para el dolor como *tipo*, pero ¿quién puede decir que no es esencial en este ejemplar particular de dolor, especialmente si este ejemplar es idéntico a un estado cerebral? Si lo es, entonces simplemente no podríamos tener el dolor particular en cuestión sin el estado cerebral particular. (Una línea argumental de este tipo es adoptada por Feldman [1974], quien sostiene que lo doloroso no es esencial en un dolor particular, y por McGinn [1977], quien sostiene que lo doloroso *y* la descarga de fibras C podrían ser esenciales para un dolor particular.) Si esto es así, entonces, al imaginar una versión incorpórea de mi dolor, no estamos imaginando *ese mismo* dolor sino un dolor separado, numéricamente distinto. Lo mismo ocurre en el caso de imaginar la descarga de mis fibras C sin que esto esté asociado al dolor. De este modo, los argumentos en contra de la identidad de ejemplares no son concluyentes, aunque el argumento en contra de la identidad de tipos puede sobrevivir.

Luego, el argumento a partir de la incorporeidad no establece un caso concluyente en contra del materialismo. Podría refutar una tesis de identidad de tipos de la clase formulada por Place y Smart, pero el

materialismo no requiere una tesis semejante.[22] Como Boyd (1980) hace notar, el materialista no necesita sostener que los estados mentales son estados físicos en todos los mundos posibles; es compatible con el materialismo que en algunos mundos los estados mentales estén constituidos por sustancia no física, en tanto que en *este* mundo están constituidos físicamente. La posibilidad de la incorporeidad sólo establece la posibilidad del dualismo, no su verdad.[23] Para ilustrar esto, podemos señalar que pocos argumentarían que la posibilidad de vida no física implica un dualismo en la biología. Es posible que todo lo que Kripke tuviese en mente fuese un argumento en contra de la tesis de identidad, pero en cualquier caso la versión más general del materialismo sobrevive.

Esto hace que el argumento a partir de la posibilidad de la instanciación de estados físicos se quede sin los correspondientes estados fenoménicos, esencialmente un argumento a partir de la posibilidad de los zombis. Curiosamente, esta es la parte del argumento de Kripke que recibió menos atención crítica, ya que la mayoría de los comentaristas se concentraron en la posibilidad de la incorporeidad. Como antes, el argumento de los zombis en contra de las tesis fuertes de identidad de tipos puede ser irrelevante, debido al hecho de que el materialismo no requiere ninguna tesis de esta clase, pero existe un argumento más general allí oculto. La posibilidad de instanciar los estados físicos relevantes sin que haya dolor, argumenta Kripke (pp. 153-54), muestra que incluso después de que Dios creó toda la sustancia física involucrada cuando tenemos un dolor —quizás un cerebro con fibras C que producen descargas— *tuvo que hacer más trabajo* para que esas descargas se experimenten como dolor. Esto es suficiente para establecer que el materialismo es falso.[24]

Este argumento de los estados físicos sin estados fenoménicos corresponde directamente al argumento que formulé en contra del materialismo. Incluso las maniobras ulteriores tienen una correspondencia. A la objeción de que esta situación es meramente concebible pero no verdaderamente posible, Kripke responderá: No podemos eliminar la situación concebida explicando que no posee la sensación de dolor pero sí el dolor mismo, ya que estar dolorido es sentir dolor en cualquier mundo posible. (Esto es, la intensión secundaria y la intensión primaria de "dolor" coinciden.) A esto podríamos agregar (con Jackson, 1980) que aun si se cuestiona la equivalencia, el argumento en contra del materialismo tendrá éxito cuando se aplica a las *sensaciones* de dolor en lugar de al propio dolor. (Esto es, aunque las intensiones difieran, el argumento todavía sigue siendo válido utilizando la intensión primaria.) Estas respuestas son

isomórficas a las que yo di para el mismo tipo de objeción anteriormente en este capítulo.

(Nótese que con su tesis de que una situación aparentemente concebible pero imposible debería eliminarse explicando que es una situación epistémicamente posible que se describe erróneamente, Kripke apoya el tratamiento "débil" de la necesidad *a posteriori*: los espacios de los mundos concebibles y posibles son uno mismo, pero los factores *a posteriori* imponen restricciones sobre su correcta descripción.[25] Para verlo, nótese que un defensor de la necesidad metafísica "fuerte", para la cual el espacio de mundos posibles es un subconjunto propio del espacio de mundos concebibles, no defendería una tesis de este tipo. En un enfoque de esta clase, podríamos describir *correctamente* una situación epistémicamente posible, pero todavía podría ser (primitivamente) metafísicamente imposible. La utilización de Kripke de la estrategia de descripción errónea, en cambio, sugiere un respaldo implícito al marco bidimensional: todos sus ejemplos de descripciones erróneas pueden verse como casos en los que se describe un mundo mediante intensiones primarias en lugar de secundarias.)

Este argumento de los estados físicos sin estados fenoménicos me parece que es la parte más convincente de la discusión de Kripke. Con frecuencia se la pasa por alto en medio de la discusión de la tesis de identidad, incorporeidad, y otras; el propio Kripke le asigna a este aspecto de su análisis un papel no central. De cualquier forma, pienso que es esta parte de la discusión la que en última instancia soporta el peso del argumento de Kripke.

En síntesis, opino que hasta donde el argumento de Kripke en contra del materialismo es exitoso, 1) la posibilidad de la incorporeidad es dudosa como argumento en contra del materialismo pero inesencial al caso; 2) los argumentos formulados en términos de identidad son igualmente dudosos pero inesenciales; 3) una metafísica esencialista es inesencial, excepto en la medida en que la sensación de dolor sea esencial al dolor como tipo, pero eso es sólo un hecho acerca de lo que "dolor" *significa*, y 4) el aparato de Kripke de designación rígida y similares no es fundamental, aunque se requiere para responder un cierto tipo de objeciones.[26] Pero este argumento contiene un núcleo sólido, en lo que es esencialmente un argumento a partir de la no superveniencia lógica.

## 4. ¿Es esto epifenomenalismo?*

Un problema, dado el enfoque que defiendo, es que si la conciencia es sólo naturalmente superveniente a lo físico, entonces parece carecer de eficacia causal. El mundo físico está más o menos

causalmente cerrado, en el sentido de que para cualquier suceso físico, parece existir una explicación física (módulo una pequeña cantidad de indeterminación cuántica). Esto implica que no hay lugar para que una conciencia no física realice algún trabajo causal independiente. Parece ser un mero epifenómeno, que depende del motor de la causalidad física, pero que no hace ninguna diferencia en el mundo físico. Existe, pero en lo que respecta al mundo físico bien podría no existir. Huxley (1874) defendió un enfoque de este tipo, pero muchas personas lo encuentran no intuitivo y repugnante. Ciertamente, esta consecuencia ha sido suficiente para hacer que algunos (por ejemplo, Kirk [1979]; Seager [1991]) cuestionen las conclusiones de sus propios argumentos en contra del materialismo, y consideren la posibilidad de que la conciencia pueda ser lógicamente superveniente a lo físico después de todo.

Este argumento ha sido formalizado en modos diferentes pero relacionados por Kirk (1979), Horgan (1987) y Seager (1991). Si suponemos que el mundo físico está causalmente cerrado y que la conciencia causa algunos sucesos físicos, entonces se deduce bajo ciertos supuestos naturales sobre la causalidad que la conciencia debe supervenir lógicamente (o metafísicamente) a lo físico.[27] Si esto es así, entonces, dado que el mundo físico está causalmente cerrado, la mera superveniencia natural de la conciencia implica que esta es epifenoménica. La forma básica del argumento es clara: si es posible sustraer lo fenoménico de nuestro mundo y todavía conservar un mundo causalmente cerrado $Z$, entonces, todo lo que ocurre en $Z$ tiene una explicación causal que es independiente de lo fenoménico, ya que no hay nada fenoménico en $Z$. Pero todo lo que ocurre en $Z$ también ocurre en nuestro mundo, de modo que la explicación causal que se aplica en $Z$ se aplica igualmente aquí. De modo que lo fenoménico es causalmente irrelevante. Aunque no exista la experiencia consciente, la conducta habría sido causada exactamente del mismo modo.

En respuesta a esto, seguiré una estrategia de dos puntas. Primero, no es obvio que la mera superveniencia natural deba implicar el epifenomenalismo en el sentido más fuerte. Es claro que el cuadro que produce *se parece algo* al epifenomenalismo. Sin embargo, la propia naturaleza de la causalidad es bastante misteriosa y es posible que cuando la comprendamos mejor estemos en una mejor posición para comprender algún modo sutil en el cual la experiencia consciente podría ser causalmente relevante. (En efecto, podría resultar que los supuestos generales en los argumentos de más arriba sean falsos.) Más adelante delinearé algunos modos como podría realizarse un análisis de este tipo. En la segunda punta de la estrategia, consideraré las *razones* de por qué el epifenomenalismo

podría resultar desagradable, y analizo su peso como argumentos. Si estas intuiciones no se traducen en argumentaciones convincentes, podría resultar que el tipo de epifenomenalismo que esta posición implica es *exclusivamente* no intuitivo, y que entonces puede aceptarse algún grado de epifenomenalismo.

## Estrategias para evitar el epifenomenalismo

Existe un número de modos en los que podríamos tratar de preservar la no superveniencia lógica pero evitando el epifenomenalismo. El más obvio de estos es negar la clausura causal de lo físico, y adoptar una forma fuerte del dualismo interaccionista en la que lo mental ocupa intersticios causales en el procesamiento físico. Pienso que esta estrategia debería evitarse, por razones que en seguida analizaré. Sin embargo, existe un número de opciones más sutiles que dependen de un enfoque apropiado de la metafísica y especialmente de la causalidad. Analizaré cuatro opciones de este tipo.

1. *Causalidad basada en la regularidad.* La primera opción es aceptar una concepción humeana fuerte de la causalidad, según la cual que A cause B significa que existe una regularidad uniforme entre sucesos de tipo A y sucesos de tipo B. Un enfoque de este tipo permitiría un papel "causal" para lo fenoménico: el mero hecho de que, por lo general, las sensaciones de dolor están seguidas por reacciones de retirada implicaría que el dolor causa estas últimas reacciones.

Una opción no humeana relacionada identifica una conexión causal con cualquier conexión *nomológica* (o legaliforme), aun cuando una regularidad nomológica es algo más que una regularidad uniforme. El enfoque de superveniencia natural es totalmente compatible con la existencia de una conexión nomológica entre la experiencia y la conducta (por ejemplo, podría haber una conexión legaliforme entre la experiencia y un estado cerebral subyacente, y una conexión legaliforme entre ese estado cerebral y la conducta). Podríamos afirmar que esto es suficiente para la causalidad. Esta aseveración podría apoyarse haciendo notar que el contrafáctico "La conducta habría sido lo misma aun en ausencia de la experiencia" es *falso* en la interpretación más natural: si la experiencia estuviese ausente, el estado cerebral habría sido diferente y la conducta habría sido diferente. Aquí, el contrafáctico se evalúa considerando mundos *naturalmente* posibles, y no mundos lógicamente posibles.

Encuentro que estas dos posiciones son poco plausibles. Ya argumenté en contra de los enfoques humeanos de la causalidad en el capítulo 2, e incluso en el enfoque no humeano no es plausible que

cualquier conexión nomológica baste para la causalidad; piénsese, por ejemplo, en la correlación entre el color del cabello de mellizos idénticos. Sin embargo, consideraciones como estas nos dan al menos una idea de por qué la conciencia *parece* tener un papel causal. Existe todo tipo de regularidades sistemáticas entre las experiencias conscientes y los sucesos físicos posteriores, cada una de los cuales nos lleva a *inferir* una conexión causal. Frente a tales regularidades, esperaríamos que las personas infieran una relación causal por razones básicamente humeanas. Esto podría, por lo tanto, eliminar algunas de nuestras intuiciones de que la conciencia es causalmente eficaz, lo que entonces apoya la segunda punta de la estrategia.

2. *Sobredeterminación causal*. Quizá podamos afirmar que un estado físico y un estado fenoménico, aunque totalmente distintos, puedan ambos causar un estado físico subsecuente. Si el estado físico $P_1$ está asociado a un estado fenoménico $Q_1$, entonces, quizá sea verdad que $P_1$ causa un estado físico posterior $P_2$ y que $Q_1$ causa $P_2$. Esto no es intuitivo: $P_1$ ya es una causa suficiente de $P_2$, así que $Q_1$ parecería ser causalmente redundante. No obstante, no es obvio que $Q_1$ no pueda estar en una relación causal con $P_1$. Esto puede ser especialmente razonable si adoptamos un enfoque no reductivo de la causalidad (del tipo propuesto por Tooley, 1987). Quizás exista una conexión causal irreducible entre los dos estados físicos y una conexión causal irreducible independiente entre el estado fenoménico y el estado físico.

Este tipo de sobredeterminación causal de sucesos suele considerarse sospechoso, pero resulta difícil mostrar de modo concluyente que haya algo malo en él. La naturaleza de la causalidad está todavía lo suficientemente mal comprendida como para que no sea posible descartar la sobredeterminación. No continuaré en esta línea, pero de todas formas merece ser tomada seriamente.

3. *La no superveniencia de la causalidad*. Una tercera estrategia se basa en la propia naturaleza de la causalidad. Vimos en el capítulo 2 que hay dos clases de hechos que no supervienen lógicamente a hechos físicos particulares: los hechos acerca de la conciencia y los hechos acerca de la causalidad. Es natural especular que estos dos resultados negativos puedan estar íntimamente relacionados, y que la conciencia y la causalidad podrían tener algún profundo vínculo metafísico. Después de todo, ambos son bastante misteriosos, y dos misterios podrían limpiamente reducirse a uno. Quizá, por ejemplo, la propia experiencia sea una especie de nexo causal; tal vez realice de algún modo la "relación causal incognoscible" de Hume; o quizá la

relación sea más compleja. Una relación como esta podría sugerir un papel para la experiencia en la causalidad que es más sutil que el tipo usual de causalidad, pero que, no obstante, evita la forma más fuerte de epifenomenalismo.

Una propuesta como esta ha sido desarrollada por Rosenberg (1996), quien sostiene que muchos de los problemas de la conciencia tienen un paralelo exacto en los problemas de la causalidad. Argumenta que debido a estos paralelos, podría ocurrir que la experiencia *realice* la causalidad, o algunos aspectos de la misma, en el mundo actual. En este enfoque, la causalidad debe realizarse en algo para sostener sus muchas propiedades, y la experiencia es un candidato natural. Si esto es así, podría ocurrir que sea la propia existencia de la experiencia lo que permite que las relaciones causales existan, de forma que la experiencia posee un tipo sutil de relevancia para la causalidad.

Por supuesto, esta propuesta es extremadamente especulativa y enfrenta algunos problemas. Para empezar, parece llevar a una versión del panpsiquismo, el enfoque de que todo es consciente, que muchos encuentran contrario a la intuición. Además, el mundo zombi es todavía un problema; pareciera que podemos imaginarnos a la causalidad ocurriendo sin la experiencia, de modo que todavía sigue pareciendo epifenoménica. Una respuesta podría ser que la causalidad debe ser realizada por algo; en el mundo zombi es realizada por alguna otra cosa, pero en este mundo la experiencia es relevante en virtud de realizar la causalidad. No me parece obvio que esta *deba* ser realizada por algo que posea alguna propiedad ulterior; si esto no es necesario, entonces la naturaleza fenoménica de la causalidad todavía seguiría siendo redundante. Pero, una vez más, la metafísica de la causalidad está lejos de ser aclarada y ciertamente vale la pena investigar esta propuesta.

*4. La naturaleza intrínseca de lo físico*. La estrategia que más me atrae surge de la observación de que la teoría física sólo caracteriza sus entidades básicas de modo *relacional*, en términos de sus relaciones causales y otras con otras entidades. Las partículas básicas, por ejemplo, se caracterizan principalmente en términos de su propensión a interactuar con otras partículas. Su masa y carga están especificadas, pero la especificación de la masa se reduce en última instancia a su propensión a ser acelerada de ciertos modos por fuerzas, etc. Cada entidad se caracteriza por su relación con otras entidades y estas se caracterizan por sus relaciones con otras entidades, y así siguiendo para siempre (excepto, quizás, por algunas entidades que se caracterizan por su relación con un observador). La

imagen del mundo físico que esto produce es la de un gigantesco flujo causal, pero la imagen no nos dice nada acerca de lo que esta causalidad *relaciona.* La referencia al protón se determina como lo que causa interacciones de un cierto tipo, que se combina en ciertos modos con otras entidades, etc.; pero, ¿qué es lo que realiza la causalidad y la combinación? Como Russell (1927) hace notar, esta es una cuestión acerca de la cual la teoría física se mantiene en silencio.[28]

Podríamos sentirnos atraídos por un enfoque del mundo como puro flujo causal, sin propiedades ulteriores que la causalidad deba relacionar, pero esto llevaría a un enfoque extrañamente insustancial del mundo físico.[29] Sólo contendría relaciones causales y nomológicas entre lugares vacíos sin características propias. Intuitivamente, es más razonable suponer que las entidades básicas que toda esta causalidad relaciona tienen alguna naturaleza interna propia, algunas propiedades *intrínsecas*, de modo que el mundo posee algo de sustancia. Sin embargo, la física puede, en el mejor de los casos, fijar la referencia de esas propiedades en virtud de sus relaciones extrínsecas; no nos dice nada directamente acerca de qué podrían ser esas propiedades. Tenemos algunas intuiciones vagas acerca de estas propiedades sobre la base de nuestra experiencia con sus análogos macroscópicos —intuiciones acerca de la "masividad" de la masa, por ejemplo— pero es difícil especificar esas intuiciones, y no es claro, si se reflexiona sobre ello, que tengan alguna importancia.

Sólo hay una clase de propiedades intrínsecas no relacionales con la que tenemos alguna familiaridad directa, y esa es la clase de las propiedades fenoménicas. Es natural especular que pueda haber alguna relación o incluso superposición entre las propiedades intrínsecas no caracterizadas de las entidades físicas y las propiedades intrínsecas familiares de la experiencia. ¿Es posible, como lo sugirió Russell, que al menos algunas de las propiedades intrínsecas de lo físico sean ellas mismas una variedad de propiedades fenoménicas?[30] La idea suena extravagante al principio, pero si se reflexiona sobre ella no lo parece tanto. Después de todo, en verdad no tenemos *ninguna idea* acerca de las propiedades intrínsecas de lo físico. Su naturaleza está abierta a cualquier propuesta, y las propiedades fenoménicas parecen ser un candidato tan probable como cualquier otro.

Por supuesto, tenemos la amenaza del panpsiquismo. No estoy seguro de que esta sea una perspectiva tan mala —si las propiedades fenoménicas son fundamentales, es natural suponer que podrían estar muy difundidas— pero no es una consecuencia necesaria. Una alternativa es que las propiedades relevantes sean propiedades protofenoménicas. En este caso la mera instanciación de una propie-

dad de esta clase no implica la experiencia, pero la instanciación de numerosas propiedades de este tipo podrían hacerlo en forma conjunta. Es difícil imaginar cómo esto podría funcionar (sabemos que no funciona para las propiedades físicas estándar), pero estas propiedades intrínsecas son bastante extrañas para nuestra concepción. La posibilidad no puede desecharse *a priori*.

De cualquier modo, este tipo de vínculo íntimo sugiere una especie de papel causal para lo fenoménico. Si existen propiedades intrínsecas en lo físico, son instanciaciones de estas propiedades las que la causalidad física relaciona. Si estas son propiedades fenoménicas, entonces existe la causalidad fenoménica; y si estas son propiedades protofenoménicas, entonces las propiedades fenoménicas heredan la relevancia causal por su condición superveniente, así como las bolas de billar heredan su relevancia causal de las moléculas. En cualquiera de los dos casos, la fenomenología de la experiencia en los agentes humanos puede heredar su relevancia causal a partir del papel causal de las propiedades intrínsecas de lo físico.

Por supuesto, esta sería una especie más sutil de relevancia causal que el tipo usual. Sigue ocurriendo, por ejemplo, que podemos imaginar que removemos las propiedades fenoménicas, y que el patrón de flujo causal permanece igual. Pero, ahora, la respuesta es que al imaginar un escenario de esta clase, estamos alterando las propiedades intrínsecas de las entidades físicas y reemplazándolas por algo más (por supuesto, el problema es que no estamos en absoluto habituados a imaginar las propiedades intrínsecas de lo físico). Así, simplemente nos movemos a un mundo en el que algo más realiza la causalidad. Si pudiese existir un mundo de *puro* flujo causal, este argumento fracasaría, pero es posible que un mundo de esta clase sea lógicamente imposible, ya que en él no habría nada que la causalidad pueda relacionar.

Esta posición es bastante parecida a la segunda posición descripta en el apartado 2, en la que los electrones tienen una esencia oculta a la cual las descripciones físicas meramente fijan su referencia. Pienso que por las razones dadas allí, las propiedades intrínsecas no deberían ser *identificadas* con propiedades físicas como la masa. Parece razonable decir que todavía hay masa en el mundo zombi, a pesar de las diferencias en su naturaleza intrínseca. Si esto es así, entonces, la masa es una propiedad extrínseca que puede "realizarse" mediante diferentes propiedades intrínsecas en diferentes mundos. Pero, cualquiera sea el modo como tomemos esta decisión semántica, la posición retiene una dualidad esencial entre las propiedades que la física trata directamente y las propiedades intrínsecas ocultas que constituyen la fenomenología.

Existe un sentido en el cual ese enfoque puede verse como un monismo en lugar de un dualismo, pero no se trata de un monismo materialista. A diferencia del fisicalismo, esta perspectiva toma ciertas propiedades fenoménicas o protofenoménicas como fundamentales. Lo que finalmente produce es una red de propiedades intrínsecas, algunas de las cuales son al menos fenoménicas o protofenoménicas y que están relacionadas de acuerdo con ciertas leyes causales/dinámicas. Estas propiedades "realizan" las propiedades físicamente extrínsecas y las leyes que las conectan realizan las leyes físicas. En el caso extremo, en el cual todas las propiedades intrínsecas son fenoménicas, podría ser mejor considerar este enfoque como una versión del idealismo. Es un idealismo muy diferente del de Berkeley, sin embargo. El mundo no es superveniente a la mente de un observador, sino que consiste en una vasta red causal de propiedades fenoménicas subyacentes a las leyes físicas postuladas por la ciencia. Un caso menos extremo en el cual las propiedades intrínsecas son protofenoménicas o en el cual algunas no son ni fenoménicas ni protofenoménicas, quizá sea mejor considerarlo como una versión del monismo neutral de Russell. Las propiedades básicas del mundo no son ni físicas ni fenoménicas, pero lo físico y lo fenoménico se construyen a partir de ellas. Lo fenoménico a partir de la combinación de sus naturalezas intrínsecas y lo físico a partir de sus relaciones extrínsecas.

Según este enfoque, las leyes más básicas serán aquellas que conectan las propiedades intrínsecas básicas. Las leyes físicas conocidas capturan la forma relacional de estas leyes, pero eliminan las propiedades intrínsecas. Las leyes psicofísicas pueden interpretarse como leyes que conectan propiedades intrínsecas (o propiedades construidas a partir de estas) con sus perfiles relacionales (o con complejas estructuras relacionales). Entonces, estas leyes no "cuelgan" ontológicamente de las leyes físicas. Más bien, las dos son consecuencias de las leyes verdaderamente básicas. Pero el orden epistemológico difiere del orden ontológico: primero somos llevados a la estructura relacional de la red causal y sólo lentamente a las propiedades intrínsecas subyacentes. Para los propósitos explicativos cotidianos, es más útil, por lo tanto, seguir pensando este enfoque en términos de una red de leyes físicas, con principios ulteriores que conectan lo físico con lo fenoménico.

Toda esta especulación metafísica debería tomarse con un grano de sal, pero muestra que la cuestión del epifenomenalismo se encuentra lejos de estar cerrada. Existe un número de cuestiones sutiles acerca de la causalidad y acerca de la naturaleza de la experiencia que

deberán ser mejor comprendidas antes de que podamos decir con seguridad si la experiencia es epifenoménica. De cualquier forma, ahora dejaré de lado la especulación metafísica y volveré a un plano menos elevado (aunque retornaré a algunas de estas cuestiones en el capítulo 8).

Sigue siendo el caso que la superveniencia natural *se siente* epifenoménica. Podríamos decir que el enfoque es epifenoménico *en una primera aproximación*: si permite alguna relevancia causal para la experiencia, lo hace de un modo sutil. Creo que podemos capturar este sentido de primera aproximación notando que el enfoque hace que la experiencia sea *explicativamente irrelevante*. Podemos dar explicaciones de la conducta en términos puramente físicos o computacionales, términos que no involucran ni implican a la fenomenología. Si la experiencia está ligada de alguna manera íntima a la causalidad, es de una forma que estas explicaciones pueden eliminar. Podríamos encontrar incluso que la irrelevancia explicativa es perturbadora; diré mucho más de esto en el próximo capítulo.

Algunos se han sentido tentados a evitar el epifenomenalismo saltando a la posición de "necesidad metafísica fuerte" del apartado 2 de este capítulo. Si la experiencia no superviene lógicamente a lo físico, el único modo, para algunos, de preservar su papel causal es declararla primitivamente idéntica o metafísicamente superveniente a alguna propiedad o propiedades físicas. Aparte de los problemas que ya he mencionado, sin embargo, el enfoque tiene además serios problemas con la irrelevancia explicativa. La conceptibilidad de un zombi muestra que, según esta perspectiva, la conducta puede explicarse en términos que no involucran ni implican la existencia de la experiencia. Las relaciones explicativas son relaciones conceptuales, de modo que la necesidad metafísica fuerte es irrelevante aquí. Este enfoque permite que la conducta sea independiente de la experiencia en un sentido fuerte, razón por la cual debe hacer frente a la mayor parte de las mismas dificultades que un dualismo de propiedades. Por lo tanto, no hay mucho que pueda ganarse al adoptar una posición de este tipo.

## ¿Dualismo interaccionista?

Algunas personas, persuadidas por los argumentos en favor del dualismo pero no convencidas de que la conciencia fenoménica deba desempeñar un papel causal significativo, podrían sentirse atraídas por una variedad interaccionista de dualismo, en la cual la experiencia llena las brechas causales en los procesos físicos. Sin embargo, ceder a esta tentación provoca más problemas de los que resuelve.

Para empezar, requiere una fuerte apuesta al futuro de la física, una que actualmente no parece en absoluto prometedora; los sucesos físicos parece que inexorablemente deben ser explicados en términos de otros sucesos físicos. También requiere una gran apuesta al futuro de la ciencia cognitiva, ya que sugiere que los tipos usuales de modelos físico/funcionales serán insuficientes para explicar la conducta. Pero, por razones que discutiré en breve, el problema más profundo es que este enfoque podría no ser mejor en evitar los problemas con el epifenomenalismo que el enfoque de clausura causal.

La única forma de dualismo interaccionista que parece aunque más no sea remotamente defendible en el panorama contemporáneo es una que hace uso de ciertas propiedades de la mecánica cuántica. Hay dos modos para esto. Primero, algunos apelaron a la existencia de la indeterminación cuántica, y sugirieron que una conciencia no física podría ser responsable de llenar las brechas causales resultantes mediante la determinación de qué valores podrían tomar algunas magnitudes físicas dentro de una distribución aparentemente "probabilística" (por ejemplo, Eccles, 1986). Aunque estas decisiones sólo tendrían un diminuto efecto proximal, quizá la dinámica no lineal podría amplificar estas diminutas fluctuaciones produciendo efectos macroscópicos significativos sobre la conducta.

Esta es una sugerencia audaz e interesante, pero tiene un número de problemas. Primero, la teoría contradice el postulado de la mecánica cuántica de que estas "decisiones" microscópicas son completamente aleatorias y, en principio, implica que debería haber algún patrón detectable en ellas, una hipótesis que es verificable. Segundo, para que esta teoría permita que la conciencia realice algún trabajo causal *interesante*, debe ocurrir que la conducta producida por *estas* decisiones microscópicas sea de algún modo diferente en especie de la producida por la mayoría de los otros conjuntos de decisiones que podrían haberse tomado mediante un proceso puramente aleatorio. Supuestamente la conducta sería más racional de lo que habría sido de otra forma, y lleva a observaciones como "Ahora estoy viendo el color rojo" que los procesos aleatorios no habrían producido. Esto, nuevamente, es verificable en principio, mediante una simulación de un cerebro con procesos aleatorios reales que determinen esas decisiones. Por supuesto no sabemos con seguridad cuál será el resultado de la verificación, pero sostener que la versión aleatoria llevaría a una conducta inusualmente degradada sería hacer una apuesta en condiciones desfavorables.

Un segundo modo en el cual la mecánica cuántica se relaciona con la cuestión de la clausura causal se encuentra en el hecho de que en algunas interpretaciones del formalismo cuántico, la propia con-

ciencia desempeña un papel causal vital, requerido para producir el llamado "colapso de la función de onda". Este colapso se supone que ocurre en todo acto de medición; y en una interpretación, el único modo de distinguir una medición de una no medición es por medio de la presencia de la conciencia. Esta teoría, ciertamente, no tiene una aceptación universal (para empezar, *presupone* que la conciencia no es ella misma física, lo que es contrario a los puntos de vista de la mayoría de los físicos) y yo mismo no la acepto, pero, en cualquier caso, parece que el tipo de trabajo causal que la conciencia realiza aquí es bastante diferente del tipo que se requiere para que la conciencia desempeñe un papel en la dirección de la conducta.[31] No es evidente de qué manera un colapso en los objetos externos percibidos permite que la conciencia afecte el procesamiento físico dentro del cerebro; las teorías de esta clase por lo general se mantienen en silencio sobre qué le ocurre al cerebro durante el colapso. E, incluso, si la conciencia de algún modo consigue colapsar el estado cerebral, entonces todas las observaciones de más arriba acerca de procesos aparentemente aleatorios y su conexión con la conducta todavía son aplicables.

En cualquier caso, todas las versiones del dualismo interaccionista tienen un problema conceptual que sugiere que son menos exitosas en evitar el epifenomenalismo de lo que podría parecer; o al menos que no están en mejores condiciones que el enfoque que propuse. Aun en estos enfoques, existe un sentido en el que lo fenoménico es irrelevante. Siempre podemos quitar el componente fenoménico de cualquier concepción explicativa, lo que produce un componente puramente causal. Imagínese (con Eccles) que existen "psicones" en la mente no física que impulsan los procesos físicos en el cerebro y que los psicones son el asiento de la experiencia. Podemos contar una historia acerca de las relaciones causales entre los psicones y los procesos físicos, y una historia acerca de la dinámica causal entre los propios psicones, sin que en ningún momento debamos invocar el hecho de que los psicones tienen propiedades fenoménicas. Igual que con los procesos físicos, podemos imaginarnos remover las propiedades *fenoménicas* de los psicones, lo que produce una situación en la que las dinámicas causales son isomórficas. De esto se desprende que el hecho de que los psicones sean el asiento de la experiencia no tiene ningún papel esencial en una explicación causal, y que aun en este cuadro la experiencia es explicativamente irrelevante.

Algunos podrían objetar que los psicones (o ectoplasma, o lo que sea) están totalmente constituidos por sus propiedades fenoménicas. Aun así, existe un sentido en el cual sus propiedades fenoménicas son irrelevantes para la explicación de la conducta; sólo son sus propieda-

des relacionales lo que importa en la historia acerca de la dinámica causal. Si objetamos que todavía tienen otras propiedades intrínsecas que son *causalmente* relevantes, tenemos una situación como la que surgió más arriba con las propiedades fenoménicas intrínsecas a las entidades *físicas.* De cualquiera de las dos formas, tenemos una especie de relevancia causal pero irrelevancia explicativa. Por cierto, no se gana nada especial apartándose de la clausura causal de lo físico. Todavía tenemos una red causal más amplia que está cerrada y sigue ocurriendo que la naturaleza fenoménica de las entidades en la red es explicativamente superflua.

Podemos incluso imaginarnos que si el interaccionismo es verdadero, entonces, por razones relativamente independientes de la experiencia consciente, nos veríamos llevados a postular eventualmente psicones para explicar la conducta, para llenar las brechas causales observadas y dar cuenta de los datos. Si esto es así, los psicones tendrían el estatus de una especie de entidad teórica como las entidades teóricas de la física. Nada en esta historia involucraría o implicaría a la experiencia, la cual sería tan superflua explicativamente como en el caso usual; podríamos contar una historia de zombis que involucre a los psicones, etc. La observación adicional de que estos psicones podrían tener propiedades fenoménicas no funciona ni mejor ni peor como respuesta al epifenomenalismo que la observación análoga de que las entidades físicas (quizá las entidades básicas o posiblemente entidades relativamente complejas) podrían tener propiedades fenoménicas además de sus características extrínsecas. La negación de la clausura causal de la física, por lo tanto, no marca una diferencia significativa en la evitación del epifenomenalismo.[32]

## Los problemas del epifenomenalismo

Cualquier enfoque que tome en serio a la conciencia deberá enfrentar al menos una forma limitada del epifenomenalismo. El hecho de que la experiencia puede ser coherentemente eliminada de cualquier concepción causal implica que la experiencia es superflua en la *explicación* de la conducta, tenga o no alguna relevancia causal sutil. Es posible que resulte ser causalmente irrelevante en un sentido más fuerte; esta cuestión está abierta. Por lo tanto, debemos seguir la segunda punta de la estrategia, y examinar exactamente cuáles son los problemas con la irrelevancia causal o explicativa de la experiencia, y determinar si en última instancia son fatales. Haré esto con mayor profundidad en el capítulo 5, pero aquí podemos examinar brevemente la cuestión.

La objeción más común al epifenomenalismo es, simplemente, que es contrario a la intuición o incluso "repugnante". Hallar que una conclusión es contraria a la intuición o repugnante no es razón *suficiente* para rechazarla, en especial si es la conclusión de una sólida argumentación. El epifenomenalismo puede ser contrario a la intuición, pero no es *obviamente* falso, de modo que si un argumento sólido nos lo impone, deberíamos aceptarlo. Por supuesto, una conclusión contraria a la intuición puede ser una buena razón para dar un paso atrás y volver a examinar el argumento, pero aún debemos encontrar algo erróneo en él sobre bases independientes. Si resulta que la falsedad de la superveniencia lógica implica el epifenomenalismo, entonces la superveniencia lógica debe ser *deseable*, pero no podemos simplemente afirmarla como si se tratase de un hecho primitivo. Para sustentar la superveniencia lógica, necesitamos alguna concepción de cómo los hechos físicos podrían implicar los hechos de la conciencia, y esto es precisamente lo que argumenté que no puede hacerse.

Las objeciones más detalladas al epifenomenalismo caen en tres clases: las que conciernen a la relación de la experiencia con la conducta ordinaria, las que conciernen a la relación de la experiencia con los *juicios* acerca de la experiencia y las que conciernen al cuadro general del mundo al que da origen.

Tómese la primera clase. Muchos encuentran sencillamente obvio que sus sensaciones de dolor causan que retiren su mano de una llama, o que mi experiencia de un dolor de cabeza no puede ser irrelevante a la explicación de por qué tomo píldoras. Ciertamente existe una fuerte intuición a este respecto. Por otro lado, podemos fácilmente eliminar explicativamente la fuente de esa intuición, en términos de las regularidades sistemáticas entre esos sucesos. Nos percatamos mucho más directamente de la experiencia y de la conducta de lo que lo hacemos de un estado cerebral subyacente; sobre la base de la exposición a regularidades sistemáticas entre la experiencia y la conducta, es natural que deba inferirse una fuerte conexión causal. Aun si la conexión sólo fuese una conexión nomológica indirecta debido a relaciones con el estado cerebral subyacente, todavía esperaríamos hacer la inferencia. De modo que esta intuición puede explicarse eliminativamente. De cualquier forma, este tipo de objeción no puede ser fatal para el enfoque, ya que es una intuición que no se extiende directamente en un argumento. Es una instancia de lo *meramente* contrario a la intuición.

La segunda clase de objeciones es más preocupante. Resulta muy extraño que nuestras experiencias deban ser irrelevantes para la explicación de por qué *hablamos* acerca de nuestras experiencias, por ejemplo, o quizás incluso para nuestros *juicios* internos acerca

de las experiencias; esto parece mucho más extraño que la mera irrelevancia de mi dolor para la explicación del retiro de mi mano. Algunos sostienen que este tipo de problema no es meramente contrario a la intuición sino fatal. Por ejemplo, podría sostenerse que esto es incompatible con nuestro *conocimiento* de la experiencia, o con nuestra capacidad para *referir* a las experiencias. Creo que cuando se detallan estos argumentos no llegan finalmente a su conclusión, pero estas preguntas son ciertamente desafiantes. Dedicaré el capítulo 5 a estas cuestiones.

Las objeciones en la tercera clase conciernen a la estructura global del enfoque. Una objeción es que la imagen es repugnante e inverosímil: la experiencia depende de lo físico mediante "colgantes nomológicos" que no están integrados a las otras leyes de la naturaleza. Pienso que esto puede combatirse desarrollando una teoría que lleve a una imagen más integrada. La etiqueta "epifenomenalismo" tiende a sugerir un enfoque en el cual la experiencia cuelga "allí arriba", de algún modo flotando libremente del procesamiento; una mejor imagen que es compatible con la superveniencia natural es una en la que la experiencia ocupa las fisuras causales. Como mínimo, podemos intentar que las leyes psicofísicas sean tan simples y elegantes como sea posible. También cae en esta clase una preocupación acerca de cómo la conciencia podría *evolucionar* en una concepción epifenomenalista, pero no es difícil ver que esto no plantea ningún problema para el enfoque que defiendo; continuaré con esta discusión al final del capítulo.

Como resultado del examen, no hay muchos *argumentos* que dañen seriamente al epifenomenalismo. La clase principal de argumentos preocupantes es la que concierne a los juicios acerca de la experiencia, que analizaré en el próximo capítulo. Argumentos aparte, algunos tienen la *intuición* de que el epifenomenalismo debe estar equivocado, pero la intuición no basta para rechazar la posición ante sólidos argumentos a su favor.

Yo no describo mi punto de vista como epifenomenalismo. La cuestión de la relevancia causal de la experiencia permanece abierta, y se necesitaría una teoría más detallada de la causalidad y de la experiencia antes de que pudiera decidirse la cuestión. Pero el enfoque implica al menos una forma débil de epifenomenalismo, y podría finalmente llevar a un epifenomenalismo de un tipo más fuerte. Aunque así ocurra, pienso que los argumentos en favor de la superveniencia natural son lo suficientemente convincentes como para que debamos aceptarlos. El epifenomenalismo es contrario a la intuición, pero las alternativas son más que contrarias a la intuición. Simplemente son *erróneas*, como ya hemos visto y volveremos a

hacerlo. La moraleja general es que si los argumentos sugieren que la superveniencia natural es verdadera, entonces deberíamos aprender a vivir con ella.

Algunos hallarán, no obstante, que la naturaleza epifenomenalista de esta posición es un error fatal. Yo tengo una cierta simpatía hacia este punto de vista, que puede considerarse como una expresión de la paradoja de la conciencia: cuando se trata de la conciencia parece que *todas* las alternativas son malas. Sin embargo, creo que los problemas de todos los otros enfoques son fatales de un modo mucho más fuerte que el hecho de que este sea contrario a la intuición. Dado que alguna opción en el espacio lógico debe ser correcta, este enfoque me parece el único candidato razonable.

## 5. La geografía lógica de las cuestiones

La argumentación de mi enfoque es una inferencia a partir de aproximadamente cuatro premisas:

1. La experiencia consciente existe.
2. La experiencia consciente no es lógicamente superveniente a lo físico.
3. Si existen fenómenos que no son lógicamente supervenientes a los hechos físicos, entonces el materialismo es falso.
4. El dominio físico está causalmente cerrado.

Las premisas 1), 2) y 3) claramente implican la falsedad del materialismo. Esto tomado en conjunción con la premisa 4) y el supuesto plausible de que seres físicamente idénticos tendrán experiencias conscientes idénticas, implica el enfoque que llamé de superveniencia natural: la experiencia consciente surge de lo físico de acuerdo con algunas leyes de la naturaleza, pero ella misma no es una entidad física. Las diversas posiciones alternativas pueden catalogarse según nieguen las premisas 1), 2), 3) o 4). Por supuesto, algunas de estas premisas pueden ser negadas de más de un modo.

Negación de la premisa 1):

i. *Eliminativismo*. En este enfoque no existen hechos positivos acerca de la experiencia consciente. Nadie está consciente en el sentido fenoménico.

Negación de la premisa 2):

La premisa 2) puede ser negada de diversos modos, dependiendo de cómo proceda la implicación en cuestión, esto es, dependiendo de

qué tipo de propiedades físicas sean centralmente responsables de implicar a la conciencia. Denomino a todo estos enfoques "materialistas reductivos", ya que todos suponen un análisis de la noción de conciencia que es compatible con la explicación reductiva.

ii. *Funcionalismo reductivo*. Este enfoque supone que la conciencia está conceptualmente implicada por lo físico en virtud de propiedades funcionales o disposicionales. En esta perspectiva, que un estado sea consciente significa que desempeña un cierto papel causal. En un mundo físicamente idéntico al nuestro, se realizarían todos los papeles causales relevantes y, por lo tanto, los estados conscientes serían todos los mismos. El mundo zombi es, entonces, lógicamente imposible.

iii. *Materialismo reductivo no funcionalista*. En este enfoque, los hechos acerca de la conciencia están conceptualmente implicados por los hechos físicos en virtud de alguna propiedad no funcional. Candidatas posibles podrían incluir propiedades bioquímicas y cuánticas, o propiedades que aún deben ser descubiertas.

iv. *Materialismo de la nueva física*. De acuerdo con este enfoque, no tenemos en la actualidad ninguna idea de cómo los hechos físicos podrían explicar a la conciencia, pero ello se debe a que nuestra concepción actual de los hechos físicos es demasiado estrecha. Cuando argumentamos que un mundo zombi es lógicamente posible, en realidad estamos argumentando que todos los campos y partículas que interactúan en el continuo espaciotemporal, postulados por la física actual, pueden existir en ausencia de la conciencia. Pero con una nueva física, las cosas podrían ser diferentes. Las entidades en un marco teórico radicalmente diferente podrían ser suficientes para implicar y explicar a la conciencia.

Negación de la premisa 3):

v. *Materialismo no reductivo*. Este es el enfoque que sostiene que aunque puede no haber una implicación lógica de los hechos físicos en los hechos acerca de la conciencia, y por lo tanto ninguna explicación reductiva de la conciencia, esta es *solamente* física. Los hechos físicos "necesitan metafísicamente" los hechos acerca de la conciencia. Aun cuando la idea de un mundo zombi es totalmente coherente, un mundo así es metafísicamente imposible.

Negación de la premisa 4):

vi. *Dualismo interaccionista*. Este enfoque acepta que la conciencia es una entidad no física, pero niega que el mundo físico esté

causalmente cerrado, de modo que la conciencia puede desempeñar un papel causal autónomo.

Luego se encuentra mi enfoque, que acepta las premisas 1), 2), 3) y 4):

vii. *Dualismo naturalista.* La conciencia superviene naturalmente a lo físico, sin supervenir lógica o "metafísicamente".

También existe un octavo enfoque común, que por lo general se encuentra subespecificado:

viii. *Materialismo no tengo idea.* "No tengo idea acerca de la conciencia. Me parece absolutamente misteriosa. Pero debe ser una entidad física, ya que el materialismo debe ser verdadero". Un enfoque de este tipo tiene una amplia difusión, pero rara vez se lo encuentra en la palabra escrita (aunque véase Fodor, 1992).

Para sintetizar rápidamente la situación tal como la veo, la opción i) parece ser manifiestamente falsa; ii) y iii) se basan en falsos análisis de la noción de la conciencia y por lo tanto cambian el tema; iv) y vi) hacen grandes e inverosímiles apuestas acerca del modo como lo físico resultará, y también tienen fatales problemas conceptuales; y v) hace una apelación inválida a la necesidad *a posteriori* kripkeana o se basa en una metafísica bizarra. Tengo una cierta simpatía por viii), pero es de suponer que eventualmente deberá reducirse a algún enfoque más específico, y ninguno de los anteriores parece funcionar. Esto deja a vii) como la única opción defendible.

Más lentamente, partiendo de las opciones iv) y vi). La opción vi), el dualismo interaccionista, requiere que lo físico tenga fisuras que puedan ser llenadas por la acción de una mente no física. La evidencia actual sugiere que esto es improbable. La opción iv) requiere que la forma de la física se transforme tan radicalmente que pueda llegar a implicar los hechos acerca de la experiencia consciente; pero nadie tiene idea de cómo *alguna* física podría hacer esto. Debido a que la física trata, en última instancia, de las propiedades estructurales y dinámicas, pareciera que todo lo que la física siempre podrá implicar es más estructura y dinámica, lo que (a menos que se acepte alguna de las otras opciones reductivas) nunca implicará la existencia de la experiencia.

La razón más profunda para rechazar las opciones iv) y vi) es que al fin de cuentas sufren del mismo problema que una física más estándar: el componente fenoménico puede ser eliminado coherentemente del componente causal. Según el enfoque interaccionista,

hemos visto que incluso si las entidades no físicas tuvieran un aspecto fenoménico, podríamos coherentemente imaginarnos que sustraemos el componente fenoménico, dejando una historia puramente causal/dinámica que caracterice la interacción y la conducta de las entidades relevantes. En el enfoque de la nueva física, aunque esta incorpore explícitamente propiedades fenoménicas, el hecho de que estas propiedades sean *fenoménicas* no puede desempeñar ningún papel esencial en la historia causal/dinámica; nos quedaríamos con una física coherente aun si eliminásemos ese aspecto. De cualquiera de los dos modos, la dinámica es todo lo que necesitamos para explicar las interacciones causales, y ningún conjunto de hechos acerca de la dinámica llega a totalizar un hecho acerca de la fenomenología. Por lo tanto todavía puede contarse una historia de zombis.

Pueden hacerse varios movimientos en respuesta, pero cada uno de estos puede también hacerse en la física estándar. Por ejemplo, quizá la dinámica abstracta falla en apreciar el hecho de que la sustancia no física en la concepción interaccionista es intrínsecamente fenoménica, de modo que las propiedades fenoménicas están profundamente involucradas en la red causal. Pero también, quizá la dinámica abstracta de la física falla en apreciar el hecho de que sus entidades básicas son intrínsecamente fenoménicas (la física las caracteriza, al fin y al cabo, sólo de un modo extrínseco), y la conclusión sería la misma. De cualquier modo, tenemos la misma clase de irrelevancia explicativa de las propiedades fenoménicas intrínsecas a la historia causal/dinámica. Por lo tanto, el movimiento hacia el interaccionismo o la nueva física no resuelven ninguno de los problemas inherentes al dualismo de propiedades que defiendo. Al fin de cuentas, puede verse que no son más que versiones más complicadas del mismo tipo de enfoque.

En lo que respecta a la opción iii), la versión más tentadora es la que se inclina por propiedades desconocidas que hasta ahora hemos pasado por alto como la clave de la implicación. Pero, finalmente, el problema es el mismo: la física sólo nos da estructura y dinámica, y estas no ascienden a una fenomenología. Las únicas propiedades disponibles parecerían ser aquellas que caracterizan la estructura física o función, o propiedades construidas a partir de estas dos. Pero, las propiedades estructurales son obviamente análisis inapropiados del concepto de experiencia y las propiedades funcionales no son mucho mejores (aunque las considero más abajo). Cualquier enfoque de este tipo, en última instancia, cambiará el tema.

Esto nos deja con las opciones i), ii), v) y vii), que corresponden a las alternativas consideradas con mayor seriedad en la literatura contemporánea: eliminativismo, funcionalismo reductivo, materia-

lismo no reductivo y dualismo de propiedades. De estas, rechazo la opción i) por estar en conflicto con los hechos manifiestos. Quizás un argumento *extraordinario* podría establecer que la experiencia consciente no existe, pero nunca he visto un argumento que se acerque ni remotamente a establecer esta cuestión. En ausencia de un argumento de este tipo, considero que la opción i) significa simplemente evadir el problema negando el fenómeno.

La opción v) suele ser atractiva para aquellos que quieren tomar en serio a la conciencia pero también conservar el materialismo. Argumenté con anterioridad que simplemente no funciona. El materialismo no reductivo defendido por Searle tiene problemas internos y colapsa en uno de los otros enfoques (probablemente en el dualismo de propiedades). Otros proponentes de este enfoque se basan en una apelación a la necesidad *a posteriori* de Kripke, pero el tipo de necesidad *a posteriori* evidenciado por ese filósofo no puede salvar al materialismo. El único modo consistente de adoptar la opción v) es recurrir a una necesidad *a posteriori* fuerte que vaya mucho más allá de la de Kripke, e invocar restricciones primitivas sobre el espacio de mundos "metafísicamente posibles". Hemos visto que no hay ninguna razón para creer en tales restricciones o creer en un tercer grado intermedio semejante de posibilidad de mundos. Esta metafísica no recibe apoyo de ningún otro fenómeno, y es difícil ver cómo *podría* ser apoyada.

Aun si se aceptase esta metafísica de la necesidad, para la mayoría de los propósitos explicativos el enfoque termina pareciéndose a la perspectiva que defiendo. Implica que la conciencia no puede ser explicada reductivamente. Implica que la experiencia consciente es explicativamente irrelevante al dominio físico. E implica que una teoría de la conciencia debe invocar principios puente que conecten los dominios físico y fenoménico, principios que no estén ellos mismos implicados en las leyes físicas. Este enfoque llama a estos principios "metafísicamente necesarios", pero para todo propósito práctico la conclusión es la misma. Este tipo de teoría tendrá la misma *forma* que las teorías dualistas que defiendo, y casi todo lo que diré al desarrollar una teoría no reductiva en los próximos capítulos se aplicará igualmente aquí.

La opción ii), el funcionalismo reductivo, es la opción materialista más seria. Dejando de lado diversas opciones extravagantes, si el materialismo es verdadero, entonces la conciencia es lógicamente superveniente, y el único modo razonable de que sea lógicamente superveniente es por medio de un análisis funcional. Según este enfoque, entonces, todo lo que *significa* que algo sea una experiencia consciente es que desempeña un cierto papel causal en un sistema.

Las propiedades fenoménicas se tratan exactamente del mismo modo que las propiedades psicológicas como el aprendizaje o la categorización.

El problema con este enfoque, por supuesto, es que representa erróneamente lo que significa ser una experiencia consciente, o ser consciente. Cuando me pregunto si otros seres son conscientes, no me estoy preguntando por sus capacidades o sus mecanismos internos, de los que ya podría saber todo; me pregunto si existe algo que es ser como ellos. Este punto puede recibir apoyo de varios modos familiares. Una forma es notar que incluso una vez que explicamos diversas capacidades funcionales, el problema de explicar la experiencia puede todavía persistir. Otra manera se apoya en la observación de que podemos imaginar que cualquier papel funcional se realice en ausencia de la experiencia consciente. Una tercera se deriva del hecho de que el conocimiento de los papeles funcionales no produce automáticamente conocimiento acerca de la conciencia. También están las objeciones, hechas anteriormente, de que un análisis funcionalista no puede dar cuenta de la determinación semántica de atribuciones de la conciencia y que funde la distinción conceptual entre conciencia y percatación.

Al fin de cuentas, el funcionalismo reductivo no difiere mucho del eliminativismo. Ambas perspectivas sostienen que existe la discriminación, la categorización, la accesibilidad, la informatividad y otros similares; y ambos niegan que haya alguna otra cosa que deba ser explicada. La diferencia principal es que la posición reductiva sostiene que algunos de estos explananda merecen el nombre de "experiencia", mientras que la posición eliminativa sostiene que ninguno de ellos lo amerita. Aparte de esta cuestión terminológica, la sustancia de los enfoques es principalmente la misma. Suele notarse que la frontera entre el reduccionismo y el eliminativismo es borrosa, y que la reducción se transforma gradualmente en eliminación cuanto más nos vemos forzados a modificar los conceptos relevantes para realizar una reducción. Al aceptar que la conciencia existe sólo en la medida en que se la define como alguna capacidad funcional, el enfoque funcionalista reductivo violenta de tal modo el concepto de conciencia que probablemente sea mejor considerarlo como una versión del eliminativismo. Ninguna de estas dos perspectivas toma en serio a la conciencia.

Esto nos deja con el enfoque vii), el dualismo de propiedades que he propuesto, como la única opción defendible. Ciertamente parece ser consecuencia de premisas bien justificadas. En cierta forma es contrario a la intuición, pero es el único enfoque que no posee un defecto fatal. Algunos encontrarán su naturaleza dualista desagradable; pero argumentaré en breve que el dualismo de esta variedad no

es irrazonable como muchos pensaron, y que está expuesto a pocas objeciones serias. Lo más inquietante de este enfoque es que implica una cierta irrelevancia de las propiedades fenoménicas en la explicación de la conducta y podría llevar al epifenomenalismo, aunque esto no ocurre automáticamente. Argumentaré en el próximo capítulo, sin embargo, que esta irrelevancia explicativa no tiene consecuencias fatales. En última instancia, este enfoque nos ofrece una perspectiva coherente, naturalista y no misteriosa de la conciencia y de su lugar en el orden natural.

## Tipo A, tipo B y tipo C

En una mirada más amplia de la geografía lógica, podemos decir que existen tres clases principales de enfoques hacia la experiencia consciente. Los enfoques de *tipo A* sostienen que la conciencia, si existe, superviene lógicamente a lo físico, por razones fundamentalmente funcionalistas o eliminativistas. Los enfoques de *tipo B* aceptan que la conciencia no es lógicamente superveniente, y sostienen que no existe ninguna implicación *a priori* de lo físico en lo fenoménico, pero de todas formas mantiene el materialismo. Los enfoques de *tipo C* niegan la superveniencia lógica y el materialismo.

Los enfoques de tipo A vienen en diversas variedades —eliminativismo, conductismo, diversas versiones del funcionalismo reductivo— pero tienen ciertas cosas en común. Un teórico de tipo A sostendrá que 1) los duplicados físicos y funcionales que carecen del tipo de experiencia que nosotros tenemos son inconcebibles; 2) María no aprende nada nuevo acerca del mundo cuando ve el color rojo por primera vez (cuanto más adquiere una destreza), y 3) todo lo que hay que explicar de la conciencia puede hacerse explicando la realización de diversas funciones. Entre los teóricos de tipo A arquetípicos se encuentran Armstrong (1968), Dennett (1991), Lewis (1966) y Ryle (1949). Otros podrían ser Dretske (1995), Rey (1982), Rosenthal (1996), Smart (1959), White (1986) y Wilkes (1984).

Los enfoques de tipo B, o versiones no reductivas del materialismo, usualmente caen presa de dificultades internas. El único enfoque de tipo B que parece ser internamente coherente es el enfoque que invoca la necesidad metafísica fuerte en un papel crucial. Adoptando este enfoque, un teórico de tipo B debe sostener que 1) los zombis y el espectro invertido son concebibles pero metafísicamente imposibles; 2) María aprende algo cuando ve el color rojo, pero este aprendizaje puede explicarse eliminativamente con un análisis al estilo de Loar, y 3) la conciencia no puede explicarse reductivamente, pero de todas formas es física. El enfoque central de tipo B nunca

recibió una formulación definitiva, pero la que está más cerca de un planteo de este tipo fue realizada por Levine (1983, 1993) y Loar (1990). Otros que parecen avalar el fisicalismo sin la superveniencia lógica son Byrne (1993), Flanagan (1992), Hill (1991), Horgan (1984b), Lycan (1995), Papineau (1993), Tye (1995) y Van Gulick (1992).

Las posiciones de tipo C incluyen diversos tipos de dualismo de propiedades, en las que el materialismo se considera falso y algún tipo de propiedades fenoménicas o protofenoménicas son consideradas irreducibles. Según este tipo de enfoques, 1) los zombis y los espectros invertidos son lógica y metafísicamente posibles, 2) María aprende algo nuevo, y su conocimiento es de hechos no físicos, y 3) la conciencia no puede explicarse reductivamente, pero podría explicarse no reductivamente en términos de leyes ulteriores de la naturaleza. Las posiciones de tipo C son adoptadas por Campbell (1970), Honderich (1981), Jackson (1982), H. Robinson (1982), W. Robinson (1988), Sprigge (1994) y en el presente trabajo.

Quizá valga la pena mencionar separadamente la posición mencionada anteriormente en la que las propiedades fenoménicas se identifican con las propiedades intrínsecas de las entidades físicas. Este tipo de enfoque es avalado por Feigl (1958), Lockwood (1989), Maxwell (1978) y Russell (1926); yo mismo tengo alguna simpatía por él. Lo incluyo como una versión de tipo C, ya que considera que las propiedades fenoménicas o protofenoménicas son fundamentales, pero tienen su propia forma metafísica. En particular, está más cerca de un monismo que la interpretación natural de tipo C. Tal vez podamos llamar a esta posición el tipo C', pero usualmente la incluiré bajo el tipo C.

Existen dos puntos de elección principales entre los tipos A, B y C. Primero, ¿es la conciencia lógicamente superveniente (tipo A versus el resto)? Segundo, ¿es el fisicalismo verdadero (tipo B versus tipo C)? Tomando el segundo punto de elección primero, no me resulta difícil rechazar el tipo B. A pesar de que tiene la virtud de tomar en serio a la conciencia, se basa en una metafísica que es incoherente u oscura, y que está básicamente inmotivada; la motivación principal es simplemente evitar el dualismo a toda costa. Al fin de cuentas, este enfoque comparte la misma forma explicativa que el tipo C, pero con una dosis agregada de misterio metafísico. En comparación, el tipo C es más directo.

La opción principal es la elección entre el tipo A y el resto. En lo que a mí respecta, el funcionalismo reductivo y el eliminativismo me parecen tan claramente falsos que encuentro difícil llegar a comprender cómo alguien podría aceptar un enfoque de tipo A. A mí me parece que sólo podemos aceptar un enfoque de ese tipo si creyésemos que no

existe ningún problema significativo acerca de la conciencia en primer lugar. Sin embargo, la experiencia indica que casi un tercio de la población está dispuesta a aceptar una posición de tipo A y no desea cambiar. Esto marca la Gran Divisoria mencionada en el prefacio: la divisoria entre los enfoques que toman a la conciencia seriamente y los que no lo hacen.

De muchos modos, la divisoria ente el tipo A y los otros es más profunda que aquella entre el tipo B y el tipo C. Esta última división involucra cuestiones relativamente sutiles de metafísica, pero la primera involucra algunas intuiciones muy básicas. Aun cuando los enfoques de tipo B y de tipo A son ambos "materialistas", los de tipo B están, en espíritu, mucho más cerca de los enfoques de tipo C. Ambas perspectivas reconocen la profundidad del problema de la conciencia mientras que los enfoques de tipo A no lo hacen.

La argumentación sólo nos puede traer hasta aquí para decidir la cuestión. Si alguien insiste en que la explicación del acceso y la informatividad explica todo, que María no descubre nada acerca del mundo cuando tiene una experiencia de rojo por primera vez y que un isomorfo funcional que sólo difiere en la experiencia consciente es inconcebible, entonces sólo puedo concluir que en lo que respecta a la experiencia estamos en planos diferentes. Quizá nuestras vidas interiores difieran notablemente. Quizás uno de nosotros está "cognitivamente cerrado" a las comprensiones del otro. Más probablemente, uno de nosotros está confundido o está a merced de un dogma. En cualquier caso, una vez que la dialéctica llega a este punto, nos hallamos ante un puente que la argumentación no puede atravesar. Más bien, hemos alcanzado una confrontación primitiva de intuiciones de una clase que es común en la discusión de las cuestiones filosóficas profundas. La argumentación explícita puede ayudarnos a aislar y caracterizar el conflicto, pero no a resolverlo.

Al comienzo de este trabajo, dije que mi enfoque adoptaba como premisa tomar en serio a la conciencia. Ahora estamos en condiciones de advertir cuáles son sus implicaciones. Tomar en serio a la conciencia es aceptar exactamente lo siguiente: que hay algo interesante que debe explicarse más allá de la realización de diversas funciones.[33] Esto tiene la condición de una premisa *prima facie* que sólo un argumento extremadamente fuerte podría cambiar. Ningún argumento que haya visto nunca se acerca a invalidar la premisa. Los teóricos de tipo A por lo general no *argumentan* en contra de la premisa, simplemente la niegan. Recíprocamente, más allá de un cierto punto es casi imposible argumentar *en favor* de la premisa, no más de lo que alguien puede argumentar que la experiencia conscien-

te existe. A lo sumo podemos tratar de explicar las cuestiones con la esperanza de que esto aporte algún esclarecimiento.

Una vez aclaradas las cuestiones, los lectores podrán decidir por sí mismos si desean tomar en serio a la conciencia. Todo lo que yo sostengo es que *si* tomamos en serio a la conciencia, entonces el dualismo de propiedades es la única opción razonable. En cuanto rechazamos el funcionalismo reductivo y el eliminativismo, se deduce inexorablemente que la conciencia no es lógicamente superveniente a lo físico. Y una vez que rechazamos la superveniencia lógica, el camino al dualismo de propiedades es inevitable. Los enfoques de tipo B son populares, pero no parecen sostenerse ante un escrutinio filosófico detallado. La principal opción metafísica que sigue abierta es aceptar un enfoque estándar de tipo C o un enfoque de tipo C'. Esta no es una cuestión que debamos decidir inmediatamente —yo mismo no tengo una opinión definida sobre ello— pero, en cualquier caso, se deduce que si queremos tomar a la conciencia en serio, debemos admitir las propiedades fenoménicas o protofenoménicas como fundamentales.

Algunos de los otros enfoques hallados en la literatura filosófica no caen explícitamente en el marco general que delineé. Habiendo construido este marco, sin embargo, no resulta difícil localizarlos y analizar sus problemas. Analizaré brevemente nueve de estas posiciones en notas al pie: el materialismo biológico,[34] el funcionalismo fisicalista,[35] el psicofuncionalismo,[36] el monismo anómalo,[37] el representacionalismo,[38] la conciencia como pensamiento de orden superior,[39] el teleofuncionalismo reductivo,[40] la causalidad emergente[41] y el misterianismo.[42]

## 6. Reflexiones sobre el dualismo naturalista

Muchas personas, incluyendo un yo mío pasado, pensaron que podían simultáneamente tomar en serio a la conciencia y seguir siendo materialistas. En este capítulo argumenté que esto no es posible, y por razones claras. La moraleja es que aquellos que quieran enfrentarse al fenómeno deben adoptar una forma de dualismo. Podríamos decir: No podemos tener la torta materialista y comer nuestra conciencia al mismo tiempo.

Aun así, muchos buscarán una alternativa a la posición que formulé, porque encuentran que su naturaleza dualista es inaceptable. Esta reacción es natural, dadas las diversas asociaciones negativas del dualismo, pero sospecho que no se basan en nada más sólido que un dogma contemporáneo. Para ver esto, vale la pena considerar

las diversas razones que podríamos tener para rechazar el dualismo en favor del materialismo, y evaluar la fuerza de estas razones tal como están las cosas.

La primera razón para preferir el materialismo es la *simplicidad*. Esta es una buena razón. Si todo el resto se mantiene igual, deberíamos preferir una teoría más simple a una que es ontológicamente libertina. La navaja de Ockham nos dice que no deberíamos multiplicar las entidades sin necesidad. Pero el resto no se mantiene igual y, en este caso, *existe* una necesidad. Hemos visto que el materialismo no puede dar cuenta de los fenómenos que deben explicarse. Así como Maxwell sacrificó una cosmovisión mecanicista simple al postular campos electromagnéticos para explicar ciertos fenómenos naturales, necesitamos sacrificar una cosmovisión fisicalista simple para explicar la conciencia. Hemos respetado debidamente a Ockham al reconocer que para derrocar al materialismo necesitamos buenos argumentos. Pero, una vez que los argumentos en contra del materialismo han sido formulados, la navaja de Ockham no puede salvarlo.

La segunda razón, y quizá la más difundida, para creer en el materialismo es inductiva: el materialismo siempre funcionó en todos lados. Para explicar fenómenos como la vida, la cognición y el clima poseemos concepciones materialistas o tenemos buenas razones para suponer que estas no están lejos. ¿Por qué la conciencia debería ser diferente?

Pero esta razón es fácil de vencer. Como hemos visto, existe una explicación simple del éxito de las concepciones materialistas en diversos dominios externos. Para fenómenos como el aprendizaje, la vida y el clima, todo lo que debe explicarse son estructuras y funciones. Debido a la clausura causal de lo físico, deberíamos esperar una concepción física de la estructura y función. Pero con la conciencia, excepcionalmente, debemos explicar más que estructuras y funciones, de modo que tenemos pocas razones para esperar que su explicación sea de un tipo similar.

Vimos en el capítulo 2 que debido a la naturaleza de nuestro acceso a los fenómenos externos, deberíamos *esperar* que una concepción materialista de cualquier fenómeno de esta clase tenga éxito. Nuestro conocimiento de estos fenómenos está físicamente mediado por la luz, el sonido y otros medios perceptuales. Dada la clausura causal de lo físico, deberíamos esperar que los fenómenos que observamos mediante estos medios sean lógicamente supervenientes a lo físico, sino nunca podríamos llegar a conocerlos. Pero nuestro acceso epistémico a la experiencia consciente es de un tipo totalmente

diferente. La conciencia está en el mismo centro de nuestro universo epistémico, y nuestro acceso a ella no está mediado por la percepción. Por lo tanto, las razones para esperar una concepción materialista de los fenómenos externos son inútiles en el caso de la conciencia, y cualquier inducción a partir de esos fenómenos será cuanto más vacilante.

Tercero, muchos prefirieron el materialismo a fin de *tomar en serio a la ciencia*. Se pensó que un enfoque dualista desafiaría a la ciencia en su propio terreno. De acuerdo con Churchland (1988), "El dualismo es inconsistente con la biología evolutiva y la física y química modernas". Pero esto es totalmente falso. Nada en el enfoque dualista que defiendo requiere que tomemos a las ciencias físicas de ningún otro modo que al pie de la letra. La clausura causal de lo físico se preserva; la física, la química, la neurociencia y la ciencia cognitiva pueden proceder como es usual. En sus propios dominios, las ciencias físicas son totalmente exitosas. Explican los fenómenos físicos de un modo admirable; simplemente fracasan en explicar la experiencia consciente.

Churchland sugiere otras razones para rechazar el dualismo: 1) la dependencia sistemática por parte de los fenómenos mentales de los fenómenos neurobiológicos; 2) los resultados computacionales modernos que sugieren que pueden lograrse resultados complejos sin un homúnculo no físico, y 3) una falta de evidencia, explicación o metodología para el dualismo. Las primeras dos razones no ofrecen ninguna evidencia en contra de mi perspectiva. Respecto de la tercera, los argumentos en favor del dualismo ya han sido presentados, mientras que la explicación y metodología dualistas se ilustrarán en el resto de esta obra.

Para muchas personas, una cuarta motivación para evitar el dualismo surge de las diversas alusiones espiritualistas, religiosas, sobrenaturales y anticientíficas de ese enfoque. Pero estas son bastante inesenciales. En la perspectiva que propongo, la conciencia está gobernada por las leyes naturales, y eventualmente podría existir una teoría científica razonable de ella. No existe ningún principio *a priori* que diga que todas las leyes naturales deben ser leyes físicas; negar el materialismo no significa negar el naturalismo. Un dualismo naturalista expande nuestro enfoque del mundo, pero no invoca a las fuerzas de la oscuridad.

En una preocupación relacionada, muchos pensaron que aceptar el dualismo sería abandonar el intento explicativo. En las palabras de Dennett (1991), "Dado el modo como el dualismo se desliza en el misterio, aceptar el dualismo es darse por vencido" (p. 37). Quizás

algunos enfoques dualistas tengan esta característica, pero esto está lejos de ser un corolario automático, como espero que el resto de esta obra deje en claro.

Ocasionalmente se escucha una quinta objeción al dualismo que consiste en que no puede explicar cómo interactúa lo físico con lo no físico. Pero la respuesta a esto es simple en el marco de la superveniencia natural: interactúan en virtud de leyes psicofísicas. Existe un sistema de leyes que asegura que una configuración física determinada estará acompañada por una experiencia particular, así como hay leyes que dictan que un objeto físico dado afectará gravitacionalmente a otros de cierto modo.

Podría objetarse que esto no nos dice cuál es la *conexión*, o *cómo* una configuración física da origen a la experiencia. Pero la búsqueda de una conexión de este tipo está desencaminada. Ni siquiera con las leyes físicas fundamentales podemos encontrar una "conexión" que haga el trabajo. Las cosas simplemente ocurren de acuerdo con la ley; más allá de un cierto punto, no se puede preguntar "cómo". Tal como Hume nos mostró, la búsqueda de conexiones últimas de este tipo es infructuosa. Si esas conexiones existen, son enteramente misteriosas tanto en los casos físico como psicofísico, de modo que el segundo no plantea aquí ningún problema especial.

Es notable que los opositores de Newton hicieron una objeción similar a su teoría de la gravitación: *¿Cómo* un cuerpo ejerce una fuerza sobre otro a la distancia? Pero la fuerza de la pregunta se disolvió con el tiempo. Hemos aprendido a vivir aceptando ciertas cosas como fundamentales.

También existe una preocupación, expresada de vez en cuando, acerca de cómo la conciencia podría haber evolucionado en un marco dualista: ¿un nuevo elemento repentinamente surgió en la naturaleza, como por arte de magia? Pero este no es un problema. Como las leyes fundamentales de la física, las leyes psicofísicas son eternas, existen desde el comienzo de los tiempos. Podría ocurrir que en las etapas iniciales del universo no hubiese nada que satisficiera los antecedentes físicos de las leyes y, por lo tanto, ninguna conciencia, aunque esto depende de la naturaleza de las leyes. De cualquier forma, con el desarrollo del universo evolucionaron ciertos sistemas físicos que satisfacían las condiciones relevantes. Cuando estos sistemas cobraron existencia, la experiencia consciente automáticamente los acompañó en virtud de las leyes en cuestión. Dado que las leyes psicofísicas existen y son intemporales, como lo sostiene el dualismo naturalista, la evolución de la conciencia no plantea ningún problema especial.

En síntesis, muy pocas de las razones usuales para rechazar el dualismo tienen alguna fuerza en contra del enfoque que propongo. La motivación residual principal para rechazar el dualismo puede simplemente encontrarse en las connotaciones negativas del término y en el hecho de que va en contra de lo que a muchos de nosotros se nos enseñó. Pero cuando miramos más allá de estas asociaciones, vemos que no hay ninguna razón para que el dualismo no sea un enfoque razonable y agradable. Ciertamente, creo que la posición que esbocé es una con la cual los que se piensan a sí mismos como materialistas, pero que toman en serio a la experiencia consciente, pueden aprender a vivir e incluso pueden llegar a apreciar.

La mía es una perspectiva que muchos que se piensan a sí mismos como "materialistas" podrían ya implícitamente compartir. Todo lo que hice fue sacar a la luz del día las implicaciones ontológicas de un enfoque natural, por ejemplo, que la conciencia "surge" de lo físico. Algunos dualistas podrían incluso encontrar que mi enfoque es demasiado materialista para su gusto, en cuyo caso que así sea. Idealmente, es un enfoque que toma lo mejor de ambos mundos y lo peor de ninguno.

Este dualismo, entonces, requiere que abandonemos poco de lo que es *importante* en nuestra actual cosmovisión científica. Solamente requiere que abandonemos un dogma. Aparte de esto, el enfoque es meramente un complemento de la cosmovisión; es una ampliación necesaria a fin de llevar la conciencia dentro de su alcance. Nuestro credo: si esto es dualismo, entonces deberíamos aprender a amar el dualismo.

# 5

# La paradoja del juicio fenoménico

## 1. Conciencia y cognición

Hasta ahora, hemos puesto de relieve las distinciones y divisiones entre la conciencia y la cognición por encima de todo lo demás. La conciencia es misteriosa; la cognición no lo es. La conciencia es ontológicamente nueva; la cognición es ontológicamente gratuita. La cognición puede explicarse funcionalmente; la conciencia se resiste a dicha explicación. La cognición está gobernada completamente por las leyes de la física; la conciencia está gobernada en parte por leyes psicofísicas independientes.

Aunque la concentración sobre estas distinciones ha sido necesaria para enfrentar las muchas y sutiles cuestiones metafísicas y explicativas que rodean a la experiencia consciente, esta puede estimular una imagen equívoca de la mente. En esta imagen, la conciencia y la cognición están completamente separadas y viven vidas independientes. Podríamos tener la impresión de que una teoría de la conciencia y una teoría de la cognición tendrán poco que ver entre sí.

Esta imagen es equívoca. Nuestra vida mental no está alienada de sí misma al modo como la imagen lo sugiere. Existen vínculos profundos y fundamentales entre la conciencia y la cognición. Por un lado, el contenido de nuestras experiencias conscientes está estrechamente relacionado con el contenido de nuestros estados cognitivos. Cada vez que tenemos una sensación de verde, individualizada fenoménicamente, tenemos una *percepción* de verde correspondiente, individualizada psicológicamente. Por otro lado, gran parte de la actividad cognitiva puede estar centrada en la experiencia

consciente. Sabemos acerca de nuestras experiencias, y emitimos juicios sobre ellas; mientras me encuentro escribiendo esto, gran parte de mi pensamientos están dedicados a la conciencia. Estas relaciones entre conciencia y cognición no son arbitrarias y caprichosas, sino sistemáticas.

Un análisis de esta relación sistemática puede proporcionar gran parte del material básico para una teoría de la conciencia. De este modo, podemos ver que la naturaleza de la cognición no es irrelevante para la conciencia, sino fundamental para su explicación. Por supuesto, una teoría de la cognición no puede hacer todo el trabajo explicativo por sí misma, pero puede desempeñar un papel importante. Después de todo, es a través de la cognición que tenemos una captación de la conciencia en primer lugar. Una investigación completa de los vínculos entre la conciencia y la cognición puede proporcionar el apoyo que necesitamos para constreñir de un modo significativo una teoría de la conciencia, y quizá lleve finalmente a una concepción de la misma que no mistifique ni trivialice los fenómenos.

En este capítulo, sentaré las bases para un estudio de la relación entre la conciencia y la cognición. El próximo apartado introduce algunas nociones que están en el centro de esa relación. El resto del capítulo es principalmente defensivo: enfrenta los diversos problemas que la relación entre la conciencia y la cognición parecerían plantear a un enfoque no reductivo. En el capítulo siguiente, comienzo la tarea de construir una teoría positiva que sistematice la relación entre la conciencia y la cognición, con el objetivo de reunirlas en un cuadro unificado de la mente.

## Los juicios fenoménicos

El nexo primario de la relación entre la conciencia y la cognición se encuentra en los *juicios fenoménicos*. Nuestra experiencia consciente no reside en un vacío fenoménico aislado. Somos conscientes de nuestra experiencia y de sus contenidos, formamos juicios acerca de ella, y nos vemos llevados a hacer aseveraciones sobre ella. Cuando tengo una sensación de rojo, a veces formo una creencia de que estoy teniendo una sensación de rojo, que puedo comunicar en un informe verbal. En un nivel más abstracto, cuando nos detenemos a reflexionar sobre los misterios que la conciencia plantea, como lo he estado haciendo a lo largo de este libro, estamos emitiendo juicios acerca de la conciencia. En un nivel más concreto, frecuentemente elaboramos juicios acerca de los *objetos* de nuestra experiencia consciente (en el ambiente, por ejemplo), como cuando pensamos, "Hay algo rojo". Denomino a los diversos juicios afines de la conciencia

*juicios fenoménicos*, no porque sean ellos mismos estados fenoménicos, sino porque se ocupan de la fenomenología o de sus objetos.

Los juicios fenoménicos suelen reflejarse en *aseveraciones* acerca de la conciencia: expresiones verbales de esos juicios. En diversos momentos, las personas hacen aseveraciones acerca de la conciencia que van desde "Tengo un dolor punzante en este momento", pasando por "El LSD me produce bizarras sensaciones de color", hasta "El problema de la conciencia es totalmente desconcertante". Estas afirmaciones y juicios están íntimamente relacionados con nuestra fenomenología, pero son en última instancia parte de nuestra psicología. Los informes verbales son actos conductuales y son, por lo tanto, susceptibles de ser explicados funcionalmente. De un modo similar los juicios fenoménicos son actos cognitivos y caen dentro del domino de la psicología.

Con frecuencia se supone que las creencias deberían considerarse estados funcionales que se caracterizan por sus vínculos causales con la conducta, el ambiente y otras creencias, pero este enfoque no tiene aceptación universal. Algunos sostienen que la experiencia fenoménica puede ser parcialmente constitutiva de la creencia o del contenido de la creencia. Para las creencias acerca de la conciencia, es probable que el enfoque funcional sea particularmente controversial: si cualquier creencia depende de la experiencia consciente, las creencias acerca de la conciencia son los candidatos más probables. Por lo tanto, adoptaré el rótulo menos cargado de "juicio" para los estados funcionales a los que me refiero, y dejaré abierta la cuestión de si un juicio acerca de la conciencia es todo lo que hay en una creencia acerca de aquella. Podemos concebir un juicio como lo que queda de una creencia después que se eliminó toda cualidad fenoménica asociada.

No debería ser una cuestión controversial que existen estados puramente psicológicos constituidos por estos juicios. Para empezar, la disposición a hacer informes verbales de una cierta forma es un estado psicológico; como mínimo podemos usar el rótulo "juicio" para esta disposición. Más aún, cuando formo una creencia acerca de mi experiencia consciente, la acompañan todo tipo de procesos funcionales, como ocurre con cualquier creencia. Estos procesos subyacen a la disposición de hacer informes verbales y a todo tipo de disposiciones. Si creemos que el LSD produce bizarras sensaciones de color, los procesos acompañantes pueden subyacer en una tendencia a permitirse o a evitar el LSD en el futuro, etc. Podemos utilizar el término "juicio" como un rótulo que cubre todos los estados o procesos que desempeñan el papel causal en cuestión. En una primera aproximación, un sistema juzga que una proposición es verdadera si tiende a responder afirmativamente cuando se le consulta sobre la proposi-

ción, si tiende a comportarse de un modo apropiado dadas sus otras creencias y deseos, etcétera.

Quizá puedan pensarse los juicios como lo que mi gemelo zombi y yo tenemos en común. Mi gemelo zombi no tiene ninguna experiencia consciente, pero *afirma* que la tiene; al menos, sus informes verbales detallados suenan como los míos. Así como yo utilizo el término, pienso que es natural decir que mi gemelo zombi *juzga* que él tiene experiencia consciente, y que sus juicios al respecto tienen una correspondencia uno a uno con los míos.

Al final de este capítulo, argumentaré que el *contenido* semántico de mis creencias fenoménicas está parcialmente constituido en formas sutiles por la propia experiencia consciente (por ejemplo, las sensaciones de rojo pueden tener un papel en constituir el contenido de ciertas creencias acerca de las sensaciones de color rojo). Si esto es así, entonces algunos de los juicios de los zombis tendrán contenidos que no son tan ricos como los contenidos correspondientes de mis creencias. No obstante, al menos corresponderán uno a uno con los míos, tendrán la misma *forma*, y *funcionarán* del mismo modo para dirigir conductas como las mías. De modo que cuando hablo del juicio de un zombi acerca de que él está teniendo una sensación de rojo, hablo acerca de *algo* interesante en su psicología: como mínimo, puede interpretarse que mis palabras refieren de un modo deflacionario al juicio que él expresa mediante las palabras "Estoy teniendo una sensación de rojo" (¡o palabras con ese sonido!). Hablaré de "aseveraciones" de un modo similar, haciendo abstracción de estas sutiles cuestiones de contenido.

Estrictamente hablando, todas las descripciones de aseveraciones y juicios fenoménicos en términos de su contenido (por ejemplo, referencias al juicio de que estamos teniendo una sensación de rojo) deberían interpretarse de este modo deflacionario. Las *creencias* fenoménicas de un sujeto seguramente poseerán el contenido completo atribuido, pero la cuestión del contenido de un juicio no es tan clara, precisamente porque no es obvio qué papel desempeña la conciencia en la constitución del contenido de una creencia fenoménica. Esta distinción no tendrá mucha utilidad en buena parte de este capítulo, ya que intentaré formular del modo más agudo posible algunos problemas que los juicios fenoménicos plantean a mi enfoque. Al final del capítulo, consideraré con más detalle estas cuestiones acerca del contenido.

## Tres clases de juicios fenoménicos

Los juicios relacionados con la experiencia consciente se clasifican por lo menos en tres grupos. Son los que llamaré juicios fenoménicos de *primer orden*, *segundo orden* y *tercer orden*. Usualmente abandonaré el calificador y hablaré de "juicios de primer orden", etc., sobreentendiendo que son siempre juicios fenoménicos.

Los juicios de *primer orden* son los juicios que acompañan a las experiencias conscientes, que conciernen no a la propia experiencia sino al *objeto* de la experiencia. Cuando tengo una sensación de rojo —cuando miro un libro rojo, por ejemplo— por lo general existe un juicio explícito o implícito, "Hay algo rojo". Cuando tengo la experiencia de oír una nota musical, existe un estado psicológico acompañante que involucra a esa nota musical. Parece apropiado decir que cualquier objeto que es *conscientemente experimentado* está también *cognitivamente representado*, aunque hay más para decir sobre esto. Junto con cada experiencia consciente existe un estado cognitivo portador de contenido. Ese estado cognitivo es lo que denomino un juicio de primer orden. (Se podría argumentar que ese estado es diferente de una creencia o un juicio de ciertos modos, como por ejemplo que no necesita ser avalado por la reflexión. Discutiré esto con mayor profundidad en el próximo capítulo, pero, por ahora, hablaré de "juicios" al menos en una primera aproximación.)

Podemos pensar que los contenidos de estos juicios de primer orden constituyen el contenido de la *percatación*, siendo esta la contraparte psicológica de la conciencia mencionada en el capítulo 1: la información de la que nos percatamos es aproximadamente la información que es accesible al sistema cognitivo, que está disponible para su informe verbal, etc. Estos juicios no son estrictamente *acerca* de la conciencia. Más bien, son *paralelos* a ella, y por lo general *acerca* de objetos y propiedades en el ambiente, o incluso en la cabeza. De hecho, es razonable decir que un juicio de primer orden es aquello sobre lo que trata la experiencia correspondiente. Cuando tengo una experiencia de un libro rojo, existe un juicio de primer orden correspondiente acerca del libro rojo. En cierto sentido, podemos decir, en consecuencia, que la experiencia y los juicios de primer orden —y por lo tanto la conciencia y la percatación— comparten sus contenidos. Formularé una concepción más refinada de esta relación en el próximo capítulo.

En el presente capítulo, me ocuparé principalmente de los juicios de *segundo orden*. Estos son más directamente juicios acerca de las experiencias conscientes. Cuando tengo una sensación de rojo, a veces noto que estoy teniendo una sensación de este tipo. Juzgo que tengo

un dolor, que experimento ciertas cualidades emocionales, etc. En general, parece que para cualquier experiencia consciente, si poseemos los recursos conceptuales relevantes, entonces tenemos al menos la capacidad para juzgar que estamos teniendo esa experiencia.

También podemos hacer juicios más detallados acerca de las experiencias conscientes. Podemos notar que estamos experimentando un matiz particularmente vivo de púrpura, o que un dolor tiene una cualidad devastadora, o incluso que una postimagen verde es la tercera de tales postimágenes que tuvimos en el día de hoy. Aparte de los juicios acerca de experiencias conscientes específicas, los juicios de segundo orden también incluyen juicios sobre *clases* particulares de experiencias conscientes, como cuando notamos que alguna droga produce sensaciones particularmente intensas o que la picazón que experimentamos antes de un estornudo es particularmente placentera.

Lo que llamo juicios de *tercer orden* son juicios acerca de la experiencia consciente como tipo. Estos van más allá de los juicios acerca de experiencias particulares. Hacemos juicios de tercer orden cuando reflexionamos sobre el hecho de que tenemos experiencias conscientes y cuando reflexionamos sobre su naturaleza. He estado haciendo juicios de tercer orden a lo largo de toda esta obra. Un típico juicio de tercer orden podría ser, "La conciencia es desconcertante; no veo como podría ser explicada reductivamente". Otros son "La experiencia consciente es inefable", e incluso "La experiencia consciente no existe".

Los juicios de tercer orden son particularmente comunes entre los filósofos, y entre aquellos que poseen una tendencia a especular sobre los misterios de la existencia. Es posible que muchas personas pasen por la vida sin hacer juicios de tercer orden. Sin embargo, dichos juicios ocurren en una clase significativa de personas. El propio hecho de que las personas realicen juicios de esa clase es algo que necesita una explicación.

Para ayudar a tener en mente estas distinciones, los diversos tipos de juicios relacionados con la conciencia pueden representarse del siguiente modo:

- Juicio de primer orden: *¡Eso es rojo!*
- Juicio de segundo orden: *Estoy teniendo una sensación de rojo en este momento.*
- Juicio de tercer orden: *Las sensaciones son misteriosas.*

## 2. La paradoja del juicio fenoménico

La existencia de los juicios fenoménicos revela una tensión fundamental dentro de una teoría no reductiva de la conciencia. El problema es el siguiente. Hemos visto que la conciencia no puede ser explicada reductivamente. Pero los juicios fenoménicos se encuentran en el dominio de la psicología y en principio deberían ser reductivamente explicables mediante los métodos usuales de la ciencia cognitiva. Debería haber una explicación física o funcional de por qué poseemos la disposición de hacer las *aseveraciones* acerca de la conciencia que hacemos, por ejemplo, y de cómo hacemos los juicios que hacemos acerca de la experiencia consciente. Se deduce entonces que nuestras aseveraciones y juicios acerca de la conciencia pueden explicarse en términos relativamente independientes de la conciencia. Dicho de un modo más enfático, parece que la conciencia es *explicativamente irrelevante* para nuestras aseveraciones y juicios acerca de la conciencia. Llamo a este resultado la paradoja del juicio fenoménico.

Esta paradoja no parece haber recibido mucha atención, pero fue formulada lúcidamente por el físico Avshalom Elitzur (1989) como un argumento en contra de los enfoques que consideran que la conciencia es "pasiva"; este autor defiende, en cambio, un dualismo interaccionista.[1] La paradoja es expresada también por el psicólogo Roger Shepard (1993), quien sugiere que es algo a lo que deberíamos resignarnos:

> En síntesis, parece que todavía nos queda un dilema: Ningún análisis de los procesos puramente físicos en un cerebro (o en un ordenador) parece capaz de capturar la cualidad particular de la experiencia subjetiva que corresponde a esos procesos. Sin embargo, un análisis de este tipo debería seguramente ser capaz de ofrecer una concepción causal de cómo un individuo llega a mecanografiar una oración como la precedente. Quizá debamos resignarnos a aceptar que aunque tanto la existencia de experiencias conscientes como las relaciones de similitud entre sus qualia tienen corporizaciones físicas con causas y efectos físicos, las experiencias conscientes o los propios qualia no son caracterizables como sucesos físicos ni comunicables entre sistemas físicos (p. 242).

Como vimos en el último capítulo, la cuestión de si la conciencia es *causalmente* irrelevante en la producción de la conducta es un tema metafísico complejo que es más conveniente dejar abierto. Pero la irrelevancia *explicativa* de la conciencia es más clara, y plantea muchas de las mismas dificultades que la irrelevancia causal. Como

sea que resulte la metafísica de la causalidad, parece relativamente evidente que puede darse una explicación física de la conducta que no recurre a, ni implica, la existencia de la conciencia.

Cuando digo en una conversación, "La conciencia es la cosa más misteriosa que existe", esto es un acto conductual. Cuando escribí en un capítulo anterior "La conciencia no puede ser explicada reductivamente", ese fue un acto conductual. Cuando comento sobre un qualia particularmente intenso de púrpura que estoy experimentando, esto es un acto conductual. Como todos los actos conductuales, estos son, en principio, explicables en términos de la organización causal interna de mi sistema cognitivo. Existe alguna historia acerca de patrones de descarga de neuronas que explicará por qué estos actos ocurrieron; en un nivel superior, probablemente haya una historia acerca de representaciones cognitivas y sus relaciones de alto nivel que hará el trabajo explicativo pertinente. Ciertamente en la actualidad no conocemos los detalles de la explicación, pero si el dominio físico es causalmente cerrado, entonces existirá alguna explicación reductiva en términos físicos o funcionales.

Cuando realizamos esta explicación de mis aseveraciones en términos físicos o funcionales, nunca estaremos obligados a invocar la existencia de la propia experiencia consciente. La explicación física o funcional se dará en forma independiente y se aplicará igualmente bien tanto a un zombi como a un experimentador consciente. Parece, por lo tanto, que la experiencia consciente es irrelevante para las explicaciones de las aseveraciones fenoménicas e irrelevante también para la explicación de los juicios fenoménicos, ¡aun cuando estas aseveraciones y juicios se ocupan fundamentalmente de la experiencia consciente!

Un modo de oponerse a esta afirmación sería argumentar que el *contenido* completo de mis aseveraciones y creencias fenoménicas no puede explicarse reductivamente porque la conciencia desempeña un papel en la constitución de ese contenido. Podríamos argumentar, por ejemplo, que las aseveraciones y creencias de un zombi son aseveraciones y creencias *diferentes* (¡aunque parecen y suenan iguales!), porque un zombi no tendría el concepto completo de conciencia. Pero, es como mínimo problemático que la conciencia deba ser irrelevante para los *sonidos* que hacemos cuando hablamos acerca de ella, a los movimientos de los dedos que estoy haciendo ahora, etc.; esta respuesta, entonces, no elimina todo el sentido de desconcierto. De manera que, por el momento, dejaré de lado esta forma de pensar acerca de las cosas, y continuaré considerando las aseveraciones y los juicios en el modo "deflacionario" que permite que puedan ser explicados reductivamente.

Otra forma de oponerse a la cuestión sería argumentar que para *cualquier* propiedad de alto nivel que pueda considerarse relevante para la explicación, existirá una explicación de bajo nivel que no invoca la existencia de esa propiedad. Podríamos argumentar que una propiedad psicológica como la memoria es explicativamente irrelevante, ya que podemos dar explicaciones neurofisiológicas de acciones que no mencionan a la memoria ni una sola vez; podríamos incluso argumentar que la temperatura es explicativamente irrelevante en la física, ya que las apelaciones explicativas a la temperatura pueden en principio reemplazarse por una concepción molecular (Kim [1989] lo denomina el problema de la *exclusión explicativa*). Esto podría sugerir que la conciencia no está en peor situación que cualquier otra propiedad de alto nivel en lo que se refiere a la irrelevancia explicativa. Si la conciencia está a la par de la memoria o la temperatura, no está en mala compañía.

Hemos visto, sin embargo, que las propiedades de alto nivel como la temperatura y la memoria son lógicamente supervenientes a lo físico. Se deduce que cuando damos una explicación de alguna acción en términos neurofisiológicos, esto no hace que la memoria sea explicativamente irrelevante. La memoria puede *heredar* la relevancia explicativa en virtud de su condición lógicamente superveniente. Cuando explicamos el deseo de compañía femenina de un hombre en términos del hecho de que es varón y no casado, ¡esto no hace que el hecho de que es soltero sea explicativamente irrelevante! El principio general aquí aplicable es que cuando dos conjuntos de propiedades están *conceptualmente* relacionados, la existencia de una explicación en términos de un conjunto no hace que el otro conjunto sea explicativamente irrelevante. En cierto sentido, una de las explicaciones puede ser una reformulación de la otra, debido a la relación conceptual entre los términos involucrados.

Cuando contamos una historia acerca de la interacción de recuerdos, existe un sentido en el cual volvemos a contar la historia física en un nivel superior de abstracción. Este nivel superior omitirá muchos detalles de la historia física y, por lo tanto, constituirá una explicación mucho más satisfactoria (todos esos detalles podrían ser parte de una confusión irrelevante), pero, no obstante, está lógicamente relacionada con la historia de nivel inferior. Lo mismo ocurre con la temperatura. Estas propiedades de alto nivel no se vuelven explicativamente irrelevantes debido a la existencia de una explicación de bajo nivel, así como la velocidad de una bola de billar no se vuelve explicativamente irrelevante debido a la existencia de procesos moleculares dentro de la bola. En general, las propiedades de alto nivel en cuestión constituirán una *redescripción* más parsimoniosa

de aquello que describe una explicación de bajo nivel. Podríamos decir que aun las descripciones de bajo nivel con frecuencia involucran *implícitamente* propiedades de alto nivel en virtud de su condición de lógicamente supervenientes, aunque no las invoquen explícitamente. Donde hay superveniencia lógica, no existe ningún problema de irrelevancia explicativa.

Los problemas con la conciencia son mucho más serios. La conciencia no es lógicamente superveniente a lo físico, de modo que no podemos sostener que una explicación física o funcional involucre implícitamente a la conciencia, o que la conciencia herede la relevancia explicativa debido a la superveniencia lógica a las propiedades involucradas en una explicación de este tipo. Una explicación física o funcional de la conducta es independiente de la conciencia en un sentido mucho más fuerte. Puede darse en términos que ni siquiera *impliquen* la existencia de la experiencia consciente. La conciencia es conceptualmente independiente de lo que entra en la explicación de nuestras aseveraciones y juicios acerca de la conciencia.

Esto no significa que *nunca* podamos apelar a la experiencia consciente en la explicación de la conducta. Es perfectamente razonable explicar el hecho de que alguien se retire de una llama haciendo notar que experimentó dolor. Después de todo, aun en el enfoque no reductivo existen regularidades legaliformes entre la experiencia y la conducta subsiguiente. Sin embargo, estas regularidades dependen en última instancia de regularidades en el nivel físico. Para cualquier explicación de la conducta que recurra a una sensación de dolor, existe una explicación más fundamental en términos puramente físicos/funcionales —quizás en términos de dolor psicológico o percepción de dolor— que no invocan o implican ninguna propiedad de la experiencia. Esta última adquiere un tipo de relevancia explicativa indirecta por su conexión nomológica con estos procesos físicos y funcionales, pero no obstante sigue siendo superflua para la explicación básica.

Para ver el problema de un modo particularmente vívido, piense en mi gemelo zombi en el universo de la puerta de al lado. Habla acerca de la experiencia consciente todo el tiempo; de hecho, parece obsesionado con ella. Emplea cantidades ridículas de tiempo encorvado sobre un ordenador, escribiendo capítulo tras capítulo sobre los misterios de la conciencia. Hace frecuentes comentarios sobre el placer que obtiene de ciertos qualia sensoriales y profesa un particular placer por los verdes y púrpuras profundos. Suele involucrarse en discusiones con zombis materialistas, argumentando que su posición no puede hacer justicia a las realidades de la experiencia consciente.

¡Y sin embargo no tiene ninguna experiencia consciente en absoluto! En su universo, los materialistas tienen la razón y él está

equivocado. La mayoría de sus aseveraciones acerca de la experiencia consciente son totalmente falsas. Pero hay ciertamente una explicación física o funcional de por qué hace las aseveraciones que hace. Después de todo, su universo está totalmente gobernado por leyes, y ningún suceso en él es milagroso, de modo que debe haber *alguna* explicación de sus aseveraciones. Pero una explicación de este tipo debe realizarse, en última instancia, en términos de procesos físicos y leyes, porque estos son los *únicos* procesos y leyes en su universo.

(Como antes, podríamos argumentar plausiblemente que un zombi no refiere a la conciencia en un sentido completo con su palabra "conciencia". Por ahora, cuando hablamos de las aseveraciones y juicios sobre la conciencia de un zombi esto debería interpretarse en el modo deflacionario discutido antes. Pero, aunque no posea el concepto completo, no hay duda de que juzga que tiene *alguna* propiedad más allá de sus propiedades estructurales y funcionales —una propiedad que llama "conciencia"— y el problema surge en esta forma con la misma fuerza.)

Mi gemelo zombi es sólo una posibilidad lógica, no una posibilidad empírica, y no deberíamos preocuparnos *demasiado* acerca de las cosas extrañas que ocurren en mundos lógicamente posibles. Sin embargo, hay motivos para sentirnos perturbados por lo que ocurre. Después de todo, cualquier explicación de la conducta de mi gemelo será también una explicación de *mi* conducta, ya que los procesos dentro de su cuerpo son precisamente reflejados por los procesos dentro del mío. La explicación de *sus* aseveraciones obviamente no depende de la existencia de la conciencia, ya que no hay conciencia en su mundo. Se deduce que la explicación de mis aseveraciones es también independiente de la existencia de la conciencia.

Para fortalecer la sensación de paradoja, nótese que mi gemelo zombi realiza un razonamiento del siguiente tipo. Se sabe que lamenta el destino de *su* gemelo zombi, que pasa todo su tiempo preocupándose por la conciencia a pesar del hecho de que no la tiene. Se preocupa acerca de lo que esto podría decir acerca de la irrelevancia explicativa de la conciencia en su propio universo. Sin embargo, sigue estando totalmente seguro de que la conciencia existe y no puede ser explicada reductivamente. Pero todo esto, para él, es un engaño monumental. No existe ninguna conciencia en su universo, en su mundo los eliminativistas siempre tuvieron la razón. A pesar del hecho de que sus mecanismos cognitivos funcionan del mismo modo que los míos, *sus* juicios acerca de la conciencia están bastante equivocados.

Esta situación paradójica es a la vez encantadora y perturbadora. No es *obviamente* fatal para la posición no reductiva, pero es, por

lo menos, algo que debemos enfrentar. Es, seguramente, la mayor tensión que debe enfrentar una teoría no reductiva, y cualquier teoría de esta clase que, como mínimo, no haga frente al problema no puede ser totalmente satisfactoria. Debemos examinar cuidadosamente las consecuencias de la situación y separar lo que es tan sólo contrario a la intuición de lo que amenaza la viabilidad de un enfoque no reductivo de la conciencia.

Nietzsche dijo: "Lo que no nos mata, nos hace más fuertes". Si podemos enfrentar esta paradoja, ello podría llevarnos a valiosas comprensiones acerca de la relación entre la conciencia y la cognición. Dedicaré el resto de este capítulo a enfrentar la paradoja, y cuestiones relacionadas acerca de la conexión entre la conciencia y la cognición se repetirán a lo largo de los próximos capítulos. De este modo, una teoría de la conciencia podrá afirmarse sobre una base mucho más sólida.

(Podríamos creer que es posible evadir la paradoja adoptando lo que he llamado una posición de tipo B, en la cual la conciencia superviene con necesidad metafísica pero no con necesidad conceptual, o una posición de tipo C', en la que las propiedades fenoménicas constituyen la naturaleza intrínseca de lo físico. Pero la paradoja surge casi tan fuertemente para estos enfoques. Aunque estas perspectivas salvan una especie de relevancia causal para la conciencia, todavía llevan a la irrelevancia explicativa, ya que la relevancia explicativa debe apoyarse en conexiones *conceptuales*. Aun en estos enfoques, podemos dar una explicación reductiva de los juicios fenoménicos pero no de la propia conciencia, lo que hace que esta sea explicativamente irrelevante para los juicios. Existirá una explicación de procesamiento de los juicios que no invoca o implica la existencia de la experiencia en ninguna etapa; la presencia de cualquier conexión "metafísicamente necesaria" ulterior o propiedades fenoménicas intrínsecas será conceptualmente independiente de cualquier cosa que forme parte de la explicación de la conducta.

Otro modo de ver esto: en estos enfoques, los zombis son todavía concebibles, y existirá una explicación perfectamente apropiada de la conducta zombi. Debido a que esta explicación se aplica a un zombi, la existencia de la conciencia no desempeñará ningún papel esencial en la misma. Pero lo que ocurre dentro de los zombis también ocurre dentro nuestro, de modo que la misma explicación se aplicará igualmente bien a nosotros. De esta forma, incluso para estas perspectivas, existirá una explicación de nuestros juicios fenoménicos para los que la conciencia es relativamente superflua.)

## Enfrentar la paradoja

Cuando se trata de la explicación de la *mayor parte* de nuestra conducta, el hecho de que la conciencia sea explicativamente irrelevante puede ser contrario a la intuición, pero no es demasiado paradójico. Para explicar que tomo el libro que está frente a mí, no necesito invocar mi *sensación* fenoménica del libro; es suficiente que invoque mi *percepción* del mismo. Cuando un oyente de un concierto suspira ante un movimiento particularmente exquisito, podríamos pensar que la cualidad experimentada de las sensaciones auditivas son fundamentales para una explicación de esa conducta, pero resulta que es posible dar una explicación totalmente en términos de percepción auditiva y respuestas funcionales. Incluso al explicar por qué retiro mi mano de la llama, será suficiente una explicación funcional en términos de la noción psicológica de dolor.

En general, resulta que allí donde podríamos creer que necesitamos invocar propiedades fenoménicas para la explicación de la conducta, usualmente podemos, en cambio, invocar propiedades psicológicas. Vimos en el capítulo 1 que existe un estado psicológico subyacente en todo estado fenoménico. Allí donde podríamos haber invocado una sensación, invocamos un registro perceptual; donde podríamos haber invocado la cualidad fenoménica de una emoción, invocamos su estado funcional correspondiente; donde podríamos haber invocado un pensamiento, sólo necesitamos invocar el contenido de ese pensamiento. Es esta correspondencia entre las propiedades fenoménicas y psicológicas lo que hace que la irrelevancia explicativa de las propiedades fenoménicas no sea un problema *demasiado* serio en general. Es contrario a la intuición al principio, pero es sólo eso. Al menos para la conducta que no está directamente involucrada en la experiencia consciente, no parece haber una necesidad urgente de invocar propiedades fenoménicas en su explicación.

Es en lo que respecta a nuestras aseveraciones y juicios sobre la conciencia que la irrelevancia explicativa de la experiencia consciente se vuelve perturbadora. Cierto, podría no ser especialmente preocupante que la conciencia sea explicativamente irrelevante para nuestros juicios fenoménicos de *primer orden*, como "Eso es una cosa roja". Es razonable que puedan explicarse exclusivamente en términos de percepción y otros procesos psicológicos; después de todo, los juicios en cuestión no están directamente involucrados con la experiencia consciente, sino con el estado del mundo. Para los juicios fenoménicos de segundo y tercer orden, sin embargo, la irrelevancia explicativa parece plantear problemas reales. Son los juicios *acerca* de la experiencia consciente y los que son responsables de que hablemos sobre

nuestras sensaciones y las preocupaciones de los filósofos acerca de los misterios de la conciencia. Una cosa es aceptar que la conciencia es irrelevante para explicar cómo camino alrededor de la habitación; otra es aceptar que es irrelevante para explicar por qué hablo acerca de la conciencia. Seguramente, nos sentimos inclinados a creer que el hecho de que soy consciente será parte de la explicación de por qué *digo* que soy consciente, o por qué *juzgo* que soy consciente; y sin embargo parece que esto no es así.

Después de todo, parte de la explicación de por qué afirmamos y juzgamos que existe el agua involucrará el hecho de que efectivamente esta existe. De un modo similar, parece que la existencia de estrellas y planetas casi seguro es explicativamente relevante para nuestro juicio de que existen estrellas y planetas. Como regla general, cuando juzgamos verdadera y fiablemente que *P*, el hecho de que *P* sea verdadero suele tener un papel fundamental en la explicación del juicio. Existen *algunos* juicios para los cuales los objetos de los mismos son explicativamente irrelevantes para los propios juicios. Piénsese en las creencias religiosas, por ejemplo, o en las creencias sobre los OVNIS, que posiblemente puedan explicarse sin invocar a dioses ni a OVNIS. Pero es bastante probable que estas sean creencias *falsas* y, obviamente, no instancias de *conocimiento*. En contraste, *sabemos* que somos conscientes.

Aquí enfrentamos una situación difícil: ¿cómo puede el conocimiento de la conciencia ser reconciliado con el hecho de que esta es explicativamente irrelevante para los juicios fenoménicos? Si los juicios fenoménicos surgen por razones independientes de la conciencia, ¿no significa esto que son injustificados? Esto, sobre todo, es la dificultad principal planteada por la paradoja del juicio fenoménico, y la analizaremos en detalle más adelante en este capítulo.

La paradoja es una consecuencia de los hechos de que 1) el dominio físico es causalmente cerrado; 2) los juicios sobre la conciencia son lógicamente supervenientes a lo físico; 3) la conciencia no es lógicamente superveniente a lo físico, y 4) sabemos que somos conscientes. De las premisas 1) y 2) se deduce que los juicios acerca de la conciencia pueden explicarse reductivamente. En combinación con la premisa 3), esto implica que la conciencia es explicativamente irrelevante para nuestros juicios, lo que está en tensión con la premisa 4). De esta forma se genera la paradoja. Podríamos tratar de escapar de ella negando alguna de sus premisas. Consideraré cada una de estas vías de escape brevemente.

Algunos dualistas negarán la premisa 1). Tradicionalmente, el dualismo interaccionista cartesiano estuvo motivado por la creencia de que sólo esto puede darle a la conciencia la relevancia respecto de

la acción que merece. Elitzur (1989) sostiene directamente a partir de la existencia de aseveraciones acerca de la conciencia la conclusión de que las leyes de la física no pueden ser completas, y de que la conciencia desempeña un papel activo en la dirección de los procesos físicos (sugiere que la segunda ley de la termodinámica podría ser falsa). Pero ya argumenté que el dualismo interaccionista es de poca ayuda para evitar el problema de la irrelevancia explicativa.

Algunos podrían sentirse tentados de negar la premisa 2), pero recuérdese que *definimos* los juicios como estados funcionales lógicamente supervenientes a lo físico. Algunos podrían objetar que no existe un estado funcional de esta clase que, aunque sólo sea remotamente, se parezca a lo que consideramos que son los juicios. Sin embargo, podemos simplemente retroceder a las *aseveraciones* acerca de la conciencia, que son actos conductuales y, por lo tanto, lógicamente supervenientes. Estas plantean las mismas dificultades que los juicios y casi tan fuertemente. Aun si alguien argumentase que los actos conductuales no son puramente físicos (se podría sostener que se requiere de la experiencia consciente para que algo pueda considerarse una *aseveración*, y no un ruido, o una aseveración acerca de la conciencia), todavía es notable que la conciencia es explicativamente irrelevante para los sonidos que producimos y para las marcas que escribimos, todos los cuales pueden interpretarse sistemáticamente como concernientes a la conciencia. De modo que surgirán problemas análogos no importa cómo definamos los estados relevantes. Sin embargo, este tipo de consideraciones puede tener, cuanto más, un papel subsidiario en el tratamiento de la paradoja, ya que es plausible que sean las *creencias* más que las *aseveraciones* las que están más estrechamente conectadas con el conocimiento, y algún tipo de contenido fenoménico de las creencias podría estar constituido por la propia experiencia. Volveré a esta cuestión más adelante en el capítulo.

Los reduccionistas y eliminativistas, por supuesto, negarán la premisa 3) o 4). He argumentado exhaustivamente en favor de 3), de modo que no repetiré aquí los argumentos. De modo similar, la negación de la premisa 4) lleva al eliminativismo, una opción que ya rechacé. Sin embargo, examinaré en breve un modo como un reduccionista podría utilizar la paradoja del juicio fenoménico.

Me parece que la actitud más razonable que podemos adoptar es reconocer que todas las premisas son probablemente verdaderas, y ver cómo pueden ser reconciliadas entre sí. Sabemos que existe la experiencia consciente; el dominio físico casi seguro es causalmente cerrado; y hemos establecido antes que la conciencia no es lógicamente superveniente a lo físico. El truco es aprender a vivir con la combinación.

## 3. Acerca de la explicación de los juicios fenoménicos

Dado lo expuesto antes, la explicación de por qué decimos las cosas que decimos acerca de la conciencia surge como un proyecto razonable e interesante para la ciencia cognitiva. Estas afirmaciones son actos conductuales, y deberían ser tan capaces de ser explicados como cualquier otro acto conductual. Debería haber ricas ganancias para cualquier científico cognitivo que siga este camino. Explicar nuestras aseveraciones y juicios acerca de la conciencia puede ser difícil, pero no será tan difícil como explicar la propia conciencia. Esta explicación no producirá automáticamente una explicación de la conciencia, por supuesto, pero bien podría señalarnos la dirección correcta.

Podemos hacer más que aceptar la posibilidad de una explicación de este tipo como conclusión intelectual, derivada de la clausura causal de la física y la superveniencia lógica de la conducta. Hay razones independientes para pensar que los juicios fenoménicos serán concomitantes naturales de ciertos tipos de procesos cognitivos, y que, si se piensa en ello, deberíamos *esperar* juicios de este tipo en un sistema inteligente con un cierto diseño. Si esto es así, entonces explicar las aseveraciones y juicios podría no ser tan difícil como pensamos; podría deducirse de algunos principios básicos acerca del diseño cognitivo.

Aquí, sólo proporcionaré un muy breve esquema de por qué podríamos creer en esto; analizaré la cuestión con mayor detalle en el capítulo 8. Para tener alguna apreciación de la situación, imagínese que hemos creado una inteligencia computacional en la forma de un agente autónomo que percibe su ambiente y tiene la capacidad de reflexionar racionalmente sobre lo que percibe. ¿Cómo debería ser un sistema de esta clase? ¿Tendría alguno concepto de conciencia o alguna noción relacionada?

Para advertir que podría tenerlo, nótese que en el diseño más natural un sistema de este tipo tendría seguramente algún concepto de sí mismo; por ejemplo, tendría la capacidad de distinguirse del resto del mundo y de otras entidades que se le parezcan. También parece razonable que un sistema de esta clase podría acceder a sus propios contenidos cognitivos mucho más directamente de lo que podría acceder a los de los otros. Si tuviese la capacidad para reflexionar, posiblemente tendría una cierta percatación directa de los propios contenidos del pensamiento, y podría razonar acerca de ese hecho. Además, un sistema de este tipo tendría muy naturalmente acceso directo a la información perceptual, de un modo similar a nuestro propio sistema cognitivo.

Si le preguntamos al sistema cómo es la percepción, ¿qué diría? ¿Diría, "No es como ninguna otra cosa"? ¿Podría decir, "Bueno, sé que hay un triciclo rojo allí, pero no tengo ni idea de *cómo* lo sé. La información sólo apareció en mi base de datos"? Quizá, pero no parece probable. Un sistema diseñado de ese modo sería bastante ineficiente y poco natural; el acceso a sus propios contenidos perceptuales será curiosamente indirecto. Parece mucho más probable que dijera, "Sé que hay un triciclo rojo porque lo *veo* allí". Cuando a su vez le preguntamos cómo sabe que está viendo el triciclo, la respuesta sería muy probablemente algo como "Solamente lo veo".

Sería un sistema extraño el que respondiese, "Sé que lo veo porque los sensores 78-84 están activados de tal y tal modo". Como Hofstadter (1979) señala, no hay ninguna necesidad de darle a un sistema un acceso tan detallado a sus partes de bajo nivel. Incluso el programa SHRDLU de Winograd (1972) no tenía conocimiento acerca del código en el que estaba escrito, a pesar del hecho de que podía percibir un mundo virtual, hacer inferencias acerca de ese mundo e incluso justificar su conocimiento en un grado limitado. Un conocimiento extra de este tipo parecería ser bastante innecesario, y sólo complicaría el proceso de percatación e inferencia.

En cambio, parece probable que un sistema de este tipo tendría el mismo tipo de actitud hacia sus contenidos perceptuales que nosotros tenemos hacia los nuestros, siendo su conocimiento de ellos directo y no mediado, al menos en lo que concierne al sistema. Cuando le preguntamos cómo sabe que ve el triciclo rojo, un sistema diseñado eficientemente diría, "¡Sólo lo *veo*!" Cuando le preguntamos cómo sabe que el triciclo es rojo, diría lo mismo que nosotros: "Sencillamente parece rojo". Si un sistema de este tipo es capaz de reflexionar, podría comenzar a preguntarse acerca de cómo es que las cosas parecen rojas, y acerca de por qué el rojo es *precisamente* un modo particular, y el azul es otro. Desde el punto de vista del sistema es sólo un hecho primitivo que el color rojo parece de un modo y el azul de otro. Por supuesto, desde nuestro punto de vista sabemos que eso se debe a que el color rojo lleva al sistema a un estado, y el azul a otro; pero desde el punto de vista de la máquina eso no ayuda.

Al reflexionar, podría comenzar a preguntarse acerca del hecho de que parece tener algún acceso a lo que está pensando y de que tiene un sentido de sí mismo. Una máquina reflexiva que fue diseñada para tener acceso directo al contenido de su percepción y pensamiento podría rápidamente comenzar a preguntarse acerca de los misterios de la conciencia (Hofstadter, 1985a, ofrece una rica discusión de esta idea): "¿Por qué el calor se *siente* de este modo?", "¿Por qué soy *yo*, y no alguien más?", "Sé que mis procesos son sólo circuitos

electrónicos, pero ¿cómo explica esto mi *experiencia* de pensamiento y percepción?".

Por supuesto, la especulación que realizo aquí no debe tomarse demasiado en serio, pero ayuda a poner de relieve la *naturalidad* del hecho de que juzgamos y afirmamos que somos conscientes, dado un diseño razonable. Sería un tipo extraño de sistema cognitivo el que no tuviese ninguna idea acerca de qué estamos hablando cuando le preguntamos cómo es ser como él. El hecho de que pensamos y hablamos acerca de la conciencia podría ser una consecuencia de características muy naturales de nuestro diseño, así como lo es en estos sistema. Y ciertamente, en la explicación de por qué estos sistemas piensan y hablan como lo hacen, nunca necesitaremos invocar una *conciencia* totalmente desarrollada. Quizás estos sistemas sean realmente conscientes y quizá no lo sean, pero la explicación es válida independientemente de este hecho. Cualquier explicación de cómo estos sistemas funcionan sólo puede realizarse en términos computacionales. En un caso de este tipo es obvio que no hay espacio para un fantasma en la máquina que desempeñe un papel explicativo.

Todo esto significa (expandiendo una aseveración del capítulo 1) que la conciencia es sorprendente, pero las afirmaciones acerca de la conciencia no lo son. Aunque la conciencia es una característica del mundo que no podemos predecir a partir de los hechos físicos, las cosas que *decimos* acerca de la misma son un fenómeno cognitivo ordinario. Alguien que supiese lo suficiente sobre la estructura cognitiva podría inmediatamente ser capaz de predecir la verosimilitud de emisiones como "Me *siento* consciente de un modo que ningún objeto físico lo estaría", o incluso el "Cogito ergo sum" de Descartes. En principio, alguna explicación reductiva en términos de procesos internos debería hacer que las aseveraciones acerca de la conciencia no sean más profundamente sorprendentes que cualquier otro aspecto de la conducta. Antes me incliné por una explicación de este tipo, y volveré a considerar la cuestión con más detalle en un capítulo posterior.

Veremos luego que los detalles de una explicación apropiada pueden ser muy útiles para lograr el despegue de una teoría de la conciencia. La relación entre una explicación de los juicios fenoménicos y una explicación de la conciencia es sutil, sin embargo. Antes de proceder, consideraré una respuesta menos sutil a la situación en la que estamos ubicados.

## ¿Es suficiente explicar los juicios?

En este punto es probable que a muchos lectores se les haya ocurrido una idea obvia, especialmente a aquellos con inclinaciones reduccionistas. Si ya hemos explicado por qué *decimos* que somos conscientes y por qué *juzgamos* que somos conscientes, ¿no explicamos ya todo lo que hay que explicar? ¿Por qué no simplemente abandonar la búsqueda de una teoría de la conciencia, y declarar que la conciencia es una quimera? Aun mejor, por qué no declarar que nuestra teoría de por qué juzgamos que somos conscientes es una teoría de la conciencia por derecho propio? Bien podría sugerirse que una teoría de nuestros juicios es toda la teoría de la conciencia que necesitamos.

Esta posición recibe algún apoyo de consideraciones acerca de juicios en otros dominios. Podría pensarse que la creencia muy difundida en dioses, que puede encontrarse en todo tipo de culturas, proporciona una excelente razón para creer que los dioses existen. Pero hay una explicación alternativa de esta creencia en términos de fuerzas sociales y psicológicas. Los ateos podrían recurrir a la inseguridad psicológica de las personas ante el cosmos, a la necesidad de una salida común para la expresión espiritual o emocional y a la naturaleza intrínsecamente autopropagante de ciertos sistemas de ideas, para explicar por qué es inevitable que las creencias religiosas se encuentren muy difundidas, dada nuestra naturaleza y circunstancias. Se puede incluso señalar la existencia de ciertos argumentos altamente plausibles, pero defectuosos, en favor de la existencia de un dios, tal como el argumento a partir del diseño y los argumentos cosmológicos. Aunque son defectuosos, no lo son *obviamente* (en particular, el argumento a partir del diseño podría razonablemente haber sido considerado convincente antes de la época de Darwin), y no es difícil ver por qué ellos deberían en general contribuir a la naturalidad de la creencia religiosa.

La observación de que la amplia difusión de las creencias religiosas podría explicarse de ese modo, sin recurrir a la existencia de dioses, suele interpretarse como evidencia de que los dioses no existen en los hechos. Según esta interpretación, la hipótesis ateísta no sólo puede explicar la estructura compleja de la naturaleza tan bien como la hipótesis teísta; también puede explicar por qué la hipótesis teísta es tan popular. Este es un modo poderoso de socavar las bases de un enfoque opositor. En el caso de la creencia religiosa, el argumento parece muy fuerte. Resulta tentador también en el caso de la conciencia.

Este es seguramente el argumento más poderoso en favor de un enfoque reductivo o eliminativista de la conciencia. Pero no es

suficiente. La analogía falla. Explicar nuestros juicios acerca de la conciencia no se acerca a eliminar sus misterios. ¿Por qué? Porque la conciencia es ella misma un explanandum. Es posible que la existencia de Dios se haya formulado como hipótesis para explicar principalmente todo tipo de hechos evidentes acerca del mundo, tales como su orden y su aparente diseño. Cuando resulta que una hipótesis alternativa puede explicar igualmente bien la evidencia, entonces no hay necesidad de la hipótesis de Dios. No existe ningún fenómeno independiente *Dios* al que podamos señalar y decir: *eso* necesita explicación. En el mejor de los casos, existe evidencia indirecta.[2] De modo similar, suele postularse que la existencia de los OVNIS explica sucesos extraños en el firmamento, marcas en el suelo, desapariciones en el Triángulo de las Bermudas, las aseveraciones de los "sobrevivientes" de los OVNIS, etc. Si resulta que esta evidencia puede explicarse sin postular la existencia de los OVNIS, entonces nuestras razones para creer en ellos desaparecen.

Pero la conciencia no es una construcción explicativa postulada para ayudar a explicar la conducta o los sucesos en el mundo. Más bien, es un explanandum primitivo, un fenómeno por derecho propio que necesita explicación. Por lo tanto, no importa si resulta que no se requiere de la conciencia para realizar ningún trabajo en la explicación de otros fenómenos. Nuestra evidencia de la conciencia nunca se encuentra en un primer plano junto con estos otros fenómenos. Aun si nuestros juicios acerca de la conciencia se explican reductivamente, *todo* lo que esto muestra es que nuestros juicios pueden explicarse reductivamente. El problema mente-cuerpo no consiste en cómo explicar nuestros juicios acerca de la conciencia. Si así fuera, sería un problema relativamente trivial. En cambio, el problema mente-cuerpo consiste en cómo explicar la propia conciencia. Si los juicios pueden explicarse sin explicar la conciencia, esto es interesante y quizá sorprendente, pero no elimina el problema mente-cuerpo.

Adoptar el enfoque de que basta con explicar nuestros juicios acerca de la conciencia (así como bastaría con explicar nuestros juicios acerca de Dios) puede entenderse naturalmente como una posición eliminativista acerca de la conciencia (así como análogamente adoptamos una posición eliminativista acerca de Dios). Como tal padece de todos los problemas que el eliminativismo típicamente enfrenta. En particular, niega la evidencia de nuestra propia experiencia. Este es el tipo de cosas que sólo puede hacer un filósofo o alguien a quien le guste enredarse en nudos intelectuales. Nuestras experiencias de rojo no desaparecen porque las neguemos. Es todavía algo que es como ser nosotros, y es todavía algo que necesita

explicación. Expulsar la propia conciencia como resultado de la paradoja del juicio fenoménico sería como arrojar al bebé con el agua de la bañera.

Existe un cierto atractivo intelectual en la posición de que basta con explicar los juicios fenoménicos. Produce la sensación de un golpe audaz que disuelve limpiamente todos los problemas y deja la confusión tirada allí en el piso frente a nosotros expuesta para que todos la vean. Sin embargo es el tipo de "solución" que sólo es satisfactoria durante medio minuto. Cuando nos detenemos a reflexionar, advertimos que todo lo que hicimos fue explicar ciertos aspectos de nuestra conducta. Explicamos por qué hablamos de cierta forma, y por qué tenemos la disposición para hacerlo así, pero ni remotamente enfrentamos el problema central, esto es, la propia experiencia consciente. Cuando terminaron los treinta segundos, nos encontramos mirando una rosa roja, inhalando su fragancia y preguntándonos: "¿Por qué la experimento de *este* modo?" Y advertimos que esta explicación no tiene nada que decir sobre la cuestión.

Si no se interpreta esta posición como una especie de eliminativismo, quizá pueda considerársela una especie de posición funcionalista, en la que la noción de conciencia se interpreta como "la entidad responsable de los juicios acerca de la conciencia". Pero esto es tan inadecuado como cualquier otra definición funcional de la conciencia. Sea o no la conciencia, *de hecho*, responsable de los juicios acerca de sí misma, esto no parece ser una verdad conceptual. Después de todo, es al menos *lógicamente* posible que podamos explicar los juicios sin explicar la conciencia, sea o no plausible; y eso es suficiente para mostrar que esta interpretación de la conciencia es falsa.

Existen otras variaciones sobre esta línea de argumentación. Por ejemplo, podríamos argumentar que existe una explicación puramente reductiva de por qué pienso que la conciencia no puede explicarse reductivamente, o de por qué pienso que la conciencia no es lógicamente superveniente, o de por qué pienso que no puede definirse funcionalmente. Podríamos incluso explicar reductivamente por qué pienso que la experiencia consciente es un explanandum. Podría pensarse que esto socava totalmente mis argumentos en los apartados anteriores y abre el camino a un enfoque reductivo de la conciencia. Pero, nuevamente, este enfoque sólo puede ser satisfactorio como una especie de esgrima intelectual. Finalmente, todavía debemos explicar por qué es *así* ser un agente consciente. Una explicación de la conducta o de algún papel causal es simplemente explicar la cosa equivocada. Esto podría parecer obstinación de mula, pero se basa en un principio simple: nuestras teorías deben explicar lo que reclama una explicación.

Esta línea de argumentación es quizá la más interesante que un reduccionista o eliminativista puede tomar —si yo fuera reduccionista, sería uno de esta clase— pero finalmente padece del problema que todas las posiciones de este tipo enfrentan: no explica lo que necesita explicación. A pesar de que esta posición es tentadora, termina fracasando en tomar en serio el problema. El problema de la conciencia no puede eliminarse de un modo tan simple.[3]

### Dennett acerca de los juicios fenoménicos

Un defensor de la posición de que nuestros juicios acerca de la conciencia es todo lo que debemos explicar es Daniel Dennett. En un artículo de 1979 escribe:

> Debemos entonces defender el enfoque de que tales juicios *agotan* nuestra conciencia inmediata, que nuestra corriente de conciencia individual consiste de nada más que tales episodios proposicionales, o mejor, que dichas corrientes de conciencia, compuestas exclusivamente de dichos episodios proposicionales, son la realidad que inspira la variedad de descripciones erróneas que pasan por teorías de la conciencia, hogareñas o académicas ... Mi punto de vista, dicho sin rodeos, es que no existe ninguna variedad fenomenológica en ninguna relación de este tipo con nuestros informes. Están los informes públicos que emitimos, y luego —en lo que a la introspección concierne— está la oscuridad. (1979, p. 95)

A esto, todo lo que puedo decir es que la introspección de Dennett es muy diferente de la mía. Cuando yo hago introspección, encuentro sensaciones, experiencias de dolor y emoción, y todo tipo de otros atavíos que, aunque *acompañados* por juicios, no *sólo* son juicios, a menos que *redefinamos* la noción de juicio, o de "episodios de nuestra percatación proposicional", para incluir tales experiencias. Si redefinimos los términos de ese modo, entonces la posición de Dennett es razonable, pero ya no hay ninguna razón para suponer que nuestros juicios pueden explicarse reductivamente. Si los juicios se interpretan, en cambio, como estados funcionalmente individualizados, tales como disposiciones a informar —como creo que es la intención de Dennett— entonces su tesis se torna poco convincente. Simplemente consiste en una negación de los datos que una teoría de la conciencia debe explicar.

¿Qué podría estar ocurriendo cuando alguien afirma que la introspección sólo revela juicios? Quizá Dennett sea un zombi.[4] Quizás él quiera decir algo inusual por "juicio". Más probablemente,

sin embargo, interpreta la introspección de otro modo: como lo que podríamos llamar *extrospección*, el proceso de observar nuestros propios mecanismos cognitivos "desde afuera", por decirlo de algún modo, y reflexionar sobre qué está ocurriendo. Observando nuestros propios *mecanismos*, es fácil llegar a la conclusión de que son los juicios los que hacen todo el trabajo. Todo lo que ocurre en los procesos cognitivos relevantes es mucha categorización, distinción y reacción. Es posible que los procesos involucrados en mi percepción de un objeto amarillo puedan explicarse completamente en términos de ciertas sensibilidades retinianas, transformaciones en representaciones internas y categorización y etiquetamiento de estas representaciones. Pero esto no explica el contenido de la introspección: sólo explica los *procesos* involucrados. La extrospección no es introspección, aunque es fácil ver cómo un filósofo inclinado a especular sobre sus propios mecanismos internos podría confundirlos. Este método explicativo no toca la experiencia consciente. (Quizá las descripciones recién hechas puedan proporcionar una excelente concepción de la fenomenología de la *visión ciega* [que se describirá en el capítulo 6], ¡pero no de la conciencia ordinaria!)

Dennett hace un movimiento similar en lo que es quizás el argumento principal de *Consciousness Explained* (1991). Luego de presentar su teoría de la informatividad, Dennett necesita argumentar que esta explica todo lo que necesita explicación y, en particular, que explica la experiencia en la medida que esta lo requiere. Después de muchas escaramuzas preliminares, formula la argumentación crucial (pp. 363-64) de que una teoría de la experiencia debe explicar por qué las cosas nos *parecen* de un cierto modo. Argumenta que su teoría puede hacerlo. Por lo tanto, concluye que su teoría explica todo lo que necesita explicación.

Este es un argumento elegante, con un timbre de plausibilidad del que carecen muchos argumentos reduccionistas acerca de la conciencia. Pero su elegancia se deriva del modo como utiliza una sutil ambigüedad en la noción de "parecer" que hace equilibrio en el filo de la navaja entre los dominios fenoménico y psicológico. Existe un sentido fenoménico de "parecer", según el cual que las cosas parezcan de un cierto modo significa que sean *experimentadas* de cierto modo. Y hay un sentido psicológico de "parecer" según el cual que las cosas parezcan de un cierto modo significa que tenemos la disposición para juzgar que son de ese modo. Es en el primer sentido que una teoría de la experiencia debe explicar el modo como las cosas parecen. Pero es en el segundo sentido que la teoría de Dennett lo explica.[5]

Una vez que se advierte este sutil equívoco, el argumento pierde la mayor parte de su fuerza. Cuando Dennett dice que su teoría

explica el modo como las cosas nos parecen, esto se reduce en última instancia a que explica por qué *decimos* que las cosas son de esa forma, y por qué nos comportamos en forma correspondiente en otros modos. (Como el propio Dennett advierte, su teoría de la conciencia se basa en su teoría cuasi conductista del contenido.) Pero esa clase de explicación no llega ni de lejos a lo que una teoría de la conciencia debe explicar. Finalmente, llamar a una teoría de este tipo una teoría de la conciencia da por sentadas todas las cuestiones importantes.

En general, cuando partimos de juicios fenoménicos como los explananda de nuestra teoría de la conciencia, inevitablemente nos veremos llevados a un enfoque reductivo. Pero los explananda últimos no son los juicios sino las propias experiencias. Ninguna mera explicación de las disposiciones conductuales explicará por qué hay algo que es como ser un agente consciente.

## 4. Argumentos en contra de la irrelevancia explicativa

Hemos visto que la paradoja de los juicios fenoménicos lleva a consecuencias contrarias a la intuición. Pero, hasta ahora, eso es todo lo que vimos. Algunas personas pensarán que las consecuencias no sólo son contrarias a la intuición sino también fatales. Para justificar esta opinión, estos objetantes necesitan un *argumento*. Un argumento de este tipo nos mostraría por qué la irrelevancia explicativa de la conciencia simplemente no puede ser verdadera.

Argumentos de este tipo son sorprendentemente difíciles de obtener, pero pueden producirse. La idea general es argumentar que la irrelevancia explicativa es inconsistente con algunos hechos bien establecidos acerca de nosotros mismos. Alcanzo a ver tres modos en lo que esto podría hacerse. Podría argumentarse que la irrelevancia explicativa es inconsistente con el hecho de que *sabemos* acerca de nuestras experiencias conscientes; o que es inconsistente con el hecho de que *recordamos* nuestras experiencias conscientes; o que es inconsistente con el hecho de que *referimos* a nuestras experiencias conscientes. No creo que ninguno de estos argumentos sea convincente, pero todos plantean cuestiones interesantes, y todos deben ser expresados.

Algunos de estos argumentos pueden formularse más naturalmente en términos de irrelevancia *causal* en lugar de irrelevancia explicativa. Con el fin de darles todo su poder, temporariamente aceptaré la irrelevancia causal de la experiencia, para ver si los argumentos tienen éxito. Es posible que puedan hacerse argumentaciones totalmente similares en términos de la irrelevancia explicativa, pero serían más complicados. De cualquier manera, ya acepté

anteriormente que *podría* resultar que la experiencia sea causalmente irrelevante, y será interesante determinar si esto podría llegar a tener consecuencias fatales.

Para tomar en consideración toda la fuerza de las objeciones de un opositor, en lo que sigue también hablaré ocasionalmente de "creencias" en lugar de "juicios". Como hice notar con anterioridad, mi línea principal de defensa no girará en torno de la distinción entre las creencias y los juicios, de modo que no le asignaré aquí mucha importancia. Sin embargo, esta cuestión podría tener todavía un papel de apoyo. En lo que sigue, el lector debería al menos mantener en el fondo de su mente que 1) cuando hablamos de creencias y juicios de zombis, estipulamos una noción deflacionaria, y 2) mis propias creencias fenoménicas, en su sentido completo, pueden estar parcialmente constituidas por la experiencia consciente.

## 5. El argumento del autoconocimiento*

El problema más difícil planteado por la irrelevancia explicativa es el que ya analicé: el conocimiento de nuestras propias experiencias conscientes. A primera vista, no sólo *juzgamos* que tenemos experiencias conscientes; *sabemos* que las tenemos. Pero si un enfoque no reductivo es correcto, entonces la experiencia es explicativamente irrelevante para la formación del juicio; el mismo juicio se habría formado aunque la experiencia estuviese ausente. Podría resultar difícil entonces ver cómo ese juicio puede considerarse *conocimiento*.

Esto podría simplemente formularse como un *desafío*: ¿Si la experiencia es explicativamente irrelevante, cómo podemos saber acerca de la experiencia? Por sí mismo, es un desafío importante, y una de las preguntas centrales acerca de la experiencia consciente. Ya existen, sin embargo, muchas preguntas difíciles, y podríamos no ser capaces de responderlas antes de desarrollar una teoría detallada de la conciencia. También puede formularse más fuertemente como un *argumento*: Si la experiencia es explicativamente irrelevante, entonces *no podríamos* saber que tenemos experiencias. Aquí me ocuparé de responder a los argumentos de este tipo. También haré algunas sugerencias en respuesta al desafío, pero ese es un proyecto que volverá a surgir más adelante.

Puedo ver dos modos relacionados en los que podría desarrollarse un argumento de este tipo. Primero, podría proceder directamente a partir de la posibilidad de mi gemelo zombi. Este hace los mismos juicios fenoménicos que yo. Cuando yo juzgo que soy consciente, él juzga que es consciente. Más aún, sus juicios se producen mediante los mismos *mecanismos* que mis juicios. Si las *justificaciones* resultan

de los juicios solamente en virtud de los mecanismos por los que se forman, como se suele suponer, entonces los juicios del zombi estarán tan justificados como los míos. Sin embargo, sus juicios no están justificados en absoluto. Después de todo, son total y sistemáticamente falsos. Parece deducirse entonces que *mis* juicios no pueden tampoco justificarse. Son producidos por los mismos mecanismos que son responsables de los juicios ilusorios de un zombi y, por lo tanto, no se los puede considerar conocimiento.

*Si* mis juicios fenoménicos no tienen más justificación que los de un zombi, entonces las bases de la posición no reductiva resultan socavadas. El propio punto inicial de la posición no reductiva, nuestro conocimiento del hecho de la experiencia, sería destruido. Se deduce que este punto funciona como un desafío y un argumento. Como desafío: ¿Cómo pueden mis juicios tener mayor justificación que los de un zombi, dado que ambos se forman a través de los mismos mecanismos? Como argumento: Si mis juicios se forman mediante los mismos mecanismos que los de un zombi, no pueden tener una mayor justificación.

El segundo argumento apela a una *teoría causal del conocimiento*. Suele sostenerse que el factor crucial en la justificación de una creencia acerca de una entidad es una conexión causal apropiada entre la creencia y la entidad de la que se trata. Mis creencias sobre la mesa que estoy mirando, por ejemplo, se justifican al menos en parte por el hecho de que la mesa es causalmente responsable de las creencias. Los proponentes de una teoría causal sostienen que un juicio acerca de algún objeto o estado de cosas debe mantener una relación causal con ese objeto o estado de cosas para que se lo pueda considerar conocimiento (quizás halla excepciones en dominios *a priori* como el del conocimiento conceptual o matemático). Ciertamente, parece que si mi creencia de que Juan está en la piscina no tiene relación causal con Juan o la piscina, entonces yo no sé que Juan está en la piscina.

Pero la experiencia es causalmente irrelevante o, por lo menos, eso acepto por ahora. Una experiencia consciente no tiene ningún papel causal en la formación de un juicio acerca de la experiencia. Si una teoría causal del conocimiento es correcta, se deduce entonces que no podemos saber nada acerca de nuestras experiencias. Nuevamente, tenemos un desafío y un argumento. El desafío: ¿Cómo puedo saber acerca de mi experiencia, dado que esta no causa mis juicios? El argumento: Si la experiencia no tiene ningún papel causal en la formación de mis juicios, entonces no se la puede considerar conocimiento.

Shoemaker (1975a) utiliza argumentos como estos para razonar en favor de un enfoque materialista de la conciencia y, de hecho, para

sostener un funcionalismo reductivo. Shoemaker supone explícitamente una teoría causal del conocimiento, argumentando que si debemos saber acerca de la experiencia, entonces esta debe causar nuestras creencias introspectivas acerca de la misma. También utiliza una versión del argumento del zombi para apoyar el funcionalismo reductivo. Si los zombis o sus equivalentes funcionales son lógicamente posibles, entonces la experiencia es inaccesible a la introspección: los zombis tienen los mismos mecanismos introspectivos que nosotros, de modo que estos mecanismos no nos permiten determinar si nosotros somos zombis o no. Shoemaker concluye que los zombis y sus equivalentes funcionales deben ser lógicamente imposibles.

Pienso que la respuesta a todos estos argumentos es bastante clara. Un dualista de propiedades debería sostener que una teoría causal del conocimiento no es apropiada para nuestro conocimiento de la conciencia, y que la justificación de nuestros juicios acerca de la misma no se encuentra en los mecanismos por los cuales esos juicios se forman. El conocimiento de la experiencia consciente es, en muchos aspectos importantes, bastante diferente del conocimiento de otros dominios. Nuestro conocimiento de la experiencia consciente no consiste en una relación causal con la experiencia, sino en algún otro tipo de relación completamente distinto.

Esta conclusión puede sustentarse sobre bases independientes. Una forma de llegar a este sustento independiente es primero considerar otro modo en el que un dualista de propiedades podría intentar responder: por medio de una teoría *fiabilista* del conocimiento. Esta podría parecer una respuesta prometedora al principio, pero creo que una teoría fiabilista es inapropiada para tratar con nuestro conocimiento de la conciencia. Resulta que una teoría causal es inapropiada por la misma razón.

En una teoría fiabilista, las creencias acerca de un tema están justificadas si se formaron mediante un proceso *fiable*; esto es, si se formaron mediante un proceso que tiende a producir creencias verdaderas. Las creencias perceptuales, por ejemplo, están justificadas si resultan de la estimulación óptica a partir de objetos en el ambiente, un proceso que por lo general produce creencias verdaderas; no están justificadas si se producen debido a alucinaciones, que son un mecanismo muy poco fiable. Es totalmente compatible con una teoría no reductiva de la experiencia que en el mundo actual, nuestros juicios fenoménicos son fiables: al menos como una cuestión de correlación nomológica, parece probable que cuando juzgamos que estamos teniendo una experiencia visual, la estamos teniendo. Los juicios fenoménicos de mi gemelo zombi, en cambio, son completamente no fiables; sus juicios son por lo general falsos.

Podría parecer, entonces, que una teoría fiabilista es la respuesta a nuestras dificultades: implica que nuestros juicios acerca de la experiencia podrían estar justificados aun en ausencia de una conexión causal directa, y tiene los recursos para explicar el hecho de que mis juicios están justificados mientras que los de mi gemelo zombi no lo están. Pero muchos podrían encontrar que la apelación a una teoría fiabilista no es, de todas formas, satisfactoria; se siente como una maniobra resbaladiza que no puede sostener la carga que se le pide que lleve. El conocimiento que una teoría fiabilista nos confiere parece demasiado débil para considerarlo el tipo de conocimiento que tenemos acerca de nuestra experiencia consciente. Si se reflexiona sobre ello, no es difícil ver por qué.

El problema es que si nuestras creencias acerca de la conciencia estuviesen justificadas *solamente* por una conexión fiable, entonces no estaríamos *seguros* de que estamos conscientes. La mera existencia de una conexión fiable no puede producir certeza, porque no tenemos ningún modo de descartar la posibilidad de que la conexión fiable esté ausente y de que no haya conciencia en el otro extremo. El único modo de estar seguros aquí sería tener algún acceso *ulterior* al otro extremo de la conexión; pero esto significaría que tenemos alguna base ulterior para nuestro conocimiento de la conciencia. Esta situación suele considerarse aceptable para nuestro conocimiento del mundo externo: no necesitamos estar seguros de que las sillas existen para *saber* (en un sentido cotidiano) que las sillas existen, de modo que no es un problema que no estemos seguros de que haya una conexión fiable entre las sillas y nuestros juicios acerca de ellas. Pero estamos seguros de que somos conscientes; al menos, esta certeza está en la base de la posición que yo defiendo. Quizás el conocimiento de que somos conscientes pueda ponerse en duda de diversos modos "filosóficos", pero no en el modo muy directo —análogo a dudar de nuestro conocimiento del mundo externo— que ocurriría si nuestras creencias estuviesen justificadas sólo por una conexión fiable.

Las creencias justificadas sólo por una conexión fiable son siempre compatibles con la existencia de *hipótesis escépticas*. Estas involucran escenarios en los que las cosas le parecen exactamente igual a un sujeto pero en los que las creencias son falsas, porque la conexión fiable no es válida. En el caso del conocimiento perceptual, por ejemplo, podemos construir una situación en el cual la conexión fiable esté ausente —una situación en la que el sujeto es un cerebro en un tanque, digamos— y todo le parecerá igual. Nada en la situación epistémica nuclear de un sujeto permite desechar este escenario. Pero en el caso de la conciencia, no podemos construir estas hipótesis escépticas. Nuestra situación epistémica nuclear ya *incluye* nuestra

experiencia consciente. No existe ninguna situación en la que todo nos parezca igual pero en la cual no estemos conscientes, ya que nuestra experiencia consciente es (por lo menos en parte) constitutiva del modo como las cosas nos parecen.

Es notable que en la construcción de escenarios escépticos relevantes para otros tipos de conocimiento, como nuestro conocimiento del mundo externo, los filósofos son siempre cuidadosos de estipular que un escenario escéptico es *experiencialmente* idéntico al escenario original. Como Descartes hizo notar, el escepticismo sólo llega hasta acá. Si un escenario escéptico involucra un conjunto vastamente diferente de experiencias en su centro —una multitud de experiencias de destellos brillantes amarillos y verdes con un sonido ensordecedor, digamos— entonces se lo desecha automáticamente. *Sabemos* (en un sentido mucho más fuerte que antes) que una situación de este tipo no es nuestra situación.

Se deduce que una concepción fiabilista del conocimiento no puede producir conocimiento que sea lo suficientemente sólido como para poseer las características de nuestro conocimiento de la experiencia consciente, y es por lo tanto inapropiado en este caso. Pero todo lo que dije acerca de una concepción fiabilista del conocimiento también se aplica a una concepción causal del mismo. Allí donde hay causalidad, hay contingencia: una conexión causal que es válida podría no haberlo sido. Si la única fuente de justificación de una creencia acerca de $X$ es una conexión causal con $X$, entonces un sujeto no puede saber con seguridad que la conexión causal exista. El único modo en el que podría saber esto con seguridad sería si tuviese algún acceso *independiente* a $X$ o a la cadena causal, pero esto implicaría conocimiento basado en algo más que la propia cadena causal. Siempre habrá un escenario escéptico en el cual todo le parece igual al sujeto, pero en el que la conexión causal está ausente y $X$ no existe; de modo que el sujeto no puede saber con seguridad acerca de $X$. Pero nosotros sabemos con seguridad que somos conscientes: de modo que una concepción causal de este conocimiento es inapropiada.

Por supuesto, un opositor podría simplemente negar que nuestro conocimiento de la conciencia sea cierto y afirmar que *existen* escenarios escépticos que no podemos desechar, un escenario zombi, por ejemplo. Pero, es probable que cualquiera que adopte este enfoque sea un eliminativista (o un funcionalista reductivo) acerca de la conciencia desde el comienzo. Si aceptamos que nuestra evidencia inmediata no descarta la posibilidad de que seamos zombis, entonces deberíamos aceptar la conclusión de que lo *somos*; para empezar, esto lleva a un enfoque mucho más simple del mundo. Pero la razón de que existe un problema con la conciencia es que nuestra evidencia inme-

diata *desecha* esa posibilidad. Tomar en serio a la conciencia es aceptar que tenemos evidencia inmediata que descarta su no existencia. Por supuesto, todo esto está abierto a discusión del modo usual; pero el punto es que no hay ninguna razón especial para comenzar a cuestionar esta cuestión en *este* punto de la argumentación. Los eliminativistas y las funcionalistas reductivos se retiraron hace mucho. *Si* tomamos en serio a la conciencia, entonces tenemos buenas razones para creer que una concepción causal o fiabilista de nuestro conocimiento fenoménico es inapropiada.

## ¿Qué justifica los juicios fenoménicos?

El problema básico con las concepciones de más arriba es que ellas hacen que nuestro acceso a la conciencia sea *mediado*, al modo como nuestro acceso a los objetos en el ambiente son mediados por algún tipo de cadena causal o mecanismo fiable. Este tipo de mediación es apropiado cuando existe una distancia entre nuestra situación epistémica nuclear y los fenómenos en cuestión, como en el caso del mundo externo: estamos conectados con objetos en el ambiente desde una cierta distancia. Pero, intuitivamente, nuestro acceso a la conciencia no está mediado en absoluto. La experiencia consciente se encuentra en el centro mismo de nuestro universo epistémico; tenemos acceso a ella *directamente*.

Esto plantea una pregunta: ¿Qué es lo que justifica las creencias acerca de nuestras experiencias, si no es un vínculo causal con esas experiencias, y si no son los mecanismos mediante los cuales se forman las creencias? Creo que la respuesta a esto es clara: es *tener* las experiencias que justifican las creencias. Por ejemplo, el propio hecho de que tengo una experiencia de rojo en este momento le proporciona justificación a mi creencia de que estoy teniendo una experiencia de rojo. Cámbiese la experiencia de rojo por un tipo diferente de experiencia, o directamente elimínesela, y la fuente principal de justificación para mi creencia habrá desaparecido. Cuando creo estar experimentando un fuerte ruido, mi garantía de esa creencia surge principalmente de mi experiencia de un ruido fuerte. Por cierto, podríamos preguntarnos, ¿de dónde más podría surgir?

Podemos señalar esta cuestión haciendo notar, como antes, que la experiencia es parte de nuestra situación epistémica nuclear. Reemplácense mis experiencias de rojo brillante por experiencias de verde mate, y cambiará mi evidencia de algunas de mis creencias, incluyendo mi creencia de que estoy teniendo una experiencia de rojo brillante. Esto se refleja en el hecho de que no hay ninguna manera de construir un escenario escéptico en el cual yo esté en una posición

epistémica cualitativamente equivalente, pero en la que mis experiencias sean radicalmente diferentes. Mis experiencias son *parte* de mi situación epistémica, y simplemente tenerlas me da evidencia de algunas de mis creencias.

Todo esto significa que hay algo intrínsecamente epistémico acerca de la experiencia. Tener una experiencia es automáticamente estar en alguna relación epistémica íntima con la misma, una relación que podríamos llamar "familiaridad". Ni siquiera hay una posibilidad conceptual de que un sujeto pueda tener una experiencia de rojo como esta sin tener *ningún* contacto epistémico con ella: tener la experiencia es estar relacionado con ella de ese modo.

Nótese que no estoy diciendo que tener una experiencia sea automáticamente *saber* acerca de ella, en el sentido en el que el conocimiento requiere creencia. Creo que esa tesis sería falsa: tenemos muchas experiencias de las que no tenemos creencias y por lo tanto no las conocemos. Más aún, podríamos tener una experiencia sin conceptualizar la experiencia de modo alguno. Tener una experiencia y, por consiguiente, estar familiarizado con ella, es estar en una relación con ella más primitiva que la creencia: proporciona evidencia a nuestras creencias, pero no constituye por sí misma una creencia.

Nada de lo que he dicho implica que todas las creencias acerca de las experiencias sean incorregibles, en el sentido que cada una de dichas creencias está totalmente justificada de modo automático. Debido a que las creencias sobre las experiencias se encuentran a una distancia de estas últimas, pueden formarse por todo tipo de razones, y a veces se formarán creencias injustificadas. Si nos distraemos, por ejemplo, podríamos hacer juicios acerca de nuestras propias experiencias que son falsos. Lo que sostengo no es que tener una experiencia sea el *único* factor que puede ser relevante para la justificación o falta de justificación de una creencia acerca de la experiencia. Lo que afirmo es simplemente que es *un* factor —quizás el factor principal— y proporciona una fuente potencial de justificación que no está presente cuando la experiencia está ausente.

A algunos les podría parecer que esto es una construcción *ad hoc* para salvar una teoría en problemas, pero yo no creo que sea *ad hoc*. Tenemos muy buenas razones, relativamente independientes de cualquier otra consideración acerca de la irrelevancia explicativa, para creer que la epistemología de la experiencia es especial, y muy diferente en especie de la epistemología de otros dominios. Muchos han hablado de nuestro "conocimiento directo" de la experiencia, o "familiaridad" con ella, sin verse forzados a adoptar esta posición como maniobra defensiva. Muchos, incluso, afirmaron que el conoci-

miento de la experiencia es la base de todo conocimiento, precisamente porque estamos en una relación epistémica directa con ella. La aseveración de que todo conocimiento se deriva del conocimiento de la experiencia puede haber sido exagerada, pero el punto general de que hay algo especial acerca de nuestro conocimiento de la experiencia nunca ha sido rebatido.[6]

De modo similar, la aseveración de que las propias experiencias justifican nuestras creencias acerca de la experiencia es fácil de motivar sobre bases independientes. Por ejemplo, en su cuidadoso análisis del conocimiento de nuestra propia mente, Siewert (1994) —quien toma en serio a la conciencia, pero que no muestra ningún signo de simpatía hacia el enfoque de que la conciencia es explicativamente irrelevante— hace una defensa en profundidad del punto de vista de que tenemos una "garantía de primera persona" sobre nuestras creencias acerca de nuestras experiencias, una garantía que se basa, al menos en parte, en que tenemos las experiencias. De modo que no es demasiado arriesgado considerar que la experiencia proporciona una fuente directa de justificación.

### Respuestas a los argumentos

Dado todo esto, la respuesta a los argumentos a partir de la irrelevancia explicativa es directa. En respuesta al argumento a partir de la teoría causal del conocimiento, notamos que existen razones independientes para creer que la teoría causal es inapropiada para explicitar nuestro conocimiento de la experiencia: nuestro conocimiento de la experiencia se basa en una relación más inmediata. Y en respuesta al argumento a partir del gemelo zombi, notamos que la justificación de mis creencias acerca de la experiencia involucran más que los mecanismos mediante los cuales las creencias se forman: crucialmente involucra a las propias experiencias. Debido a que mi gemelo zombi carece de experiencias, está en una situación epistémica muy diferente de mí mismo, y sus juicios carecen de la justificación correspondiente.

Podría ser tentador objetar que si mi creencia se encuentra en el dominio físico, su justificación debe encontrarse en dicho dominio; sin embargo esta es una conclusión errónea. A partir del hecho de que no existe ninguna justificación en el dominio físico, podríamos concluir que la porción *física* de mí (mi cerebro, digamos) no está justificada en su creencia. Pero la cuestión es si *yo* estoy justificado en la creencia, no si mi *cerebro* lo está, y si el dualismo de propiedades es correcto entonces hay más en mí que mi cerebro. Yo estoy constituido por propiedades físicas y no físicas, y la historia completa acerca de mí no

puede contarse concentrándonos sólo en una mitad. En particular, la justificación de mi creencia resulta no sólo en virtud de mis características físicas sino también en virtud de algunas de mis características no físicas, particularmente las propias experiencias.

Todavía podría objetarse, "¡Pero la creencia se habría formado de todos modos aunque la experiencia hubiese estado ausente!". A esto respondo, "¿Y qué?". En *este* caso, tengo *evidencia* de mi creencia, particularmente mi inmediata familiaridad con la experiencia. En una situación diferente, esa evidencia estaría ausente. Notar que en un caso diferente la creencia podría haberse formado en ausencia de la evidencia no significa que la evidencia no justifique la creencia en ese caso.[7] Yo *sé* que soy consciente, y el conocimiento se basa solamente en mi experiencia inmediata. Decir que la experiencia no hace ninguna diferencia en mi funcionamiento psicológico no significa que la experiencia no haga ninguna diferencia en *mí*.

Finalmente, existe un persistente estribillo que surge en estas situaciones: "¡Pero su gemelo zombi diría la misma cosa!". Si yo digo que sé que soy consciente, debe advertirse que mi gemelo zombi dice lo mismo. Si yo digo que mi creencia está justificada por mi familiaridad inmediata con la experiencia, se advierte que mi gemelo zombi dice lo mismo. A esto, la respuesta es nuevamente, "¿Y qué?". Cuanto más, esto muestra que desde el punto de vista de *tercera persona*, mi gemelo zombi y yo somos idénticos, de modo que *usted* no puede estar seguro de que yo sea consciente; pero esto siempre lo supimos. Sin embargo, no hace nada por implicar que desde el enfoque de *primera persona*, yo no puedo saber que soy consciente. Desde el punto de vista de primera persona, mi gemelo zombi y yo somos muy diferentes: yo tengo experiencias y él no. A pesar del hecho de que él dice las mismas cosas que yo, yo sé que no soy él (aunque *usted* podría no estar seguro) debido a mi familiaridad directa de primera persona con mis experiencias. Esto podría sonar algo paradójico al principio, pero en realidad significa simplemente lo obvio: nuestra experiencia de la conciencia nos permite saber que somos conscientes.

Aun cuando se objete que mi gemelo zombi *creería* las mismas cosas que yo, esto no contribuye en absoluto a hacer plausible la hipótesis escéptica de primera persona de que yo podría ser un zombi. Subyacente en este tipo de objeción podría estar el supuesto implícito de que las propias creencias son los determinantes principales de mi situación epistémica; de modo que si hay una situación en la cual yo creo exactamente las mismas cosas que sé ahora, es una situación que es evidencialmente equivalente a la mía actual. Pero, por supuesto, esto es falso. La evidencia de mis creencias acerca de las experiencias es mucho más primitiva que las propias creencias. Es la

experiencia la que es primaria; las creencias son fundamentalmente un fenómeno secundario.

También, debería recordarse que estamos estipulando una noción *deflacionaria* (esto es, funcional) de la creencia, de modo que decir que mi gemelo zombi cree las mismas cosas que yo sólo significa hacer una aseveración acerca de nuestros puntos en común desde el punto de vista de tercera persona: él está dispuesto a hacer los mismos tipos de aseveraciones, los mismos tipos de inferencias, etc. Esto no dice nada acerca de cómo son las cosas desde el interior. La idea de que "un zombi tendría las mismas creencias" constituye una objeción en este punto podría surgir de suponer una noción *inflacionaria* de la creencia, según la cual esta sería, al menos en parte, un fenómeno experiencial. Sólo en ese sentido podría ocurrir que la identidad en las creencias haga que las situaciones sean indistinguibles desde el punto de vista de primera persona; pero, por supuesto, en ese sentido no hay ninguna razón para aceptar que un zombi tenga las mismas creencias en primer lugar.

El resultado de todo esto es que los argumentos acerca del autoconocimiento no proporcionan ninguna razón para rechazar el enfoque que defiendo. Si tomamos en serio a la conciencia, tendremos buenas razones para adoptar una epistemología de la conciencia que haga que esos argumentos se queden sin dientes. Aunque existen muchas argumentaciones tentadoras que pueden hacerse, ninguna de ellas parece resistir el escrutinio.

Por supuesto, queda mucho por hacer para clarificar la epistemología de primera persona de la conciencia. Lo más que he hecho es un escueto bosquejo de un marco teórico para pensar acerca de estas cosas; quedan muchas cuestiones por tratar. En particular, nos gustaría un análisis de exactamente cómo una experiencia justifica una creencia; qué otros factores son relevantes para justificar las creencias acerca de las experiencias; bajo precisamente qué circunstancias una creencia acerca de la experiencia está completamente justificada; etc. Todas estas son cuestiones importantes que merecen ser consideradas en detalle en un estudio de la epistemología de la conciencia. Pero todas son parte del *desafío*, y a primera vista no hay razones para creer que el desafío no pueda ser enfrentado. Lo importante es que los *argumentos* que parecería que el autoconocimiento proporciona en contra de un enfoque no reductivo de la experiencia no tienen éxito.

## 6. El argumento a partir de la memoria*

La segunda objeción a la irrelevancia causal o explicativa de la experiencia es que es incompatible con el hecho de que *recordamos* nuestras experiencias. Ciertamente, parece que suelo rememorar mis antiguas experiencias, como cuando recuerdo el olor picante de la naftalina en un armario cuando era niño, o cuando recuerdo una experiencia particularmente vívida de naranja mientras observaba ayer la puesta del sol. Pero recordar algo, suele sostenerse, es estar en una relación causal apropiada con ello; esto se conoce a veces como la *teoría causal de la memoria*. Sin embargo, si las experiencias son causalmente irrelevantes para mi funcionamiento psicológico, entonces parece que mis antiguas experiencias no están causalmente relacionadas con ninguno de mis estados actuales. Si esto fuera así, entonces no podríamos recordar nuestras experiencias en absoluto.

La teoría causal de la memoria no está escrita en piedra, sin embargo. Surge de un análisis de lo que parece más apropiado de decir acerca de diversas situaciones. Como en el caso del conocimiento, puede ocurrir que una teoría causal sea apropiada en muchos dominios sin que sea apropiada en todos ellos. En particular, no es obvio que sea adecuada en el dominio de la experiencia. Las teorías causales podrían no ser *tan* inapropiadas en el caso de la memoria como lo son en el caso del conocimiento, ya que no hay ninguna duda de que nuestra relación con una experiencia recordada está mediatizada, y es plausible que gran parte de esa mediación involucre una cadena causal. Pero esto no significa que la cadena causal cuente la historia completa.

En el caso de las experiencias recordadas, ciertamente existirá una conexión causal en el nivel de la *psicología*: el estado cognitivo subyacente en el momento de la experiencia original estará causalmente conectado con el estado cognitivo en el momento del recuerdo. Y parece plausible que una conexión causal apropiada de este tipo es todo lo que se requiere para el recuerdo de la experiencia. Por ejemplo, puede existir una conexión causal entre una *creencia* fenoménica en un momento anterior y creencias en un momento posterior; y si lo que dije en el apartado previo es correcto, esta creencia original puede ser considerada *conocimiento*, ya que está justificada por una familiaridad con la propia experiencia. Este tipo de conexión causal entre una creencia justificada por la familiaridad y una creencia posterior parece relativamente suficiente para que la creencia posterior pueda considerarse una instancia de recuerdo. De modo que parecen existir buenas razones para creer que no se requiere una conexión causal con una experiencia para recordarla.

Por supuesto, la cuestión de exactamente qué puede considerarse un "recuerdo" y qué meramente una creencia verdadera justificada acerca del pasado es fundamentalmente una decisión semántica en estos casos. Lo importante es que una teoría no reductiva puede salvar las apariencias suministrando un mecanismo mediante el cual se forman las creencias verdaderas acerca de las experiencias pasadas. En la medida que una teoría no reductiva puede hacer esto, entonces cualquier *argumento* que la memoria proporciona en contra de una teoría de este tipo se queda sin colmillos. Si alguien *insiste* en que es necesario para el recuerdo una conexión causal con un objeto, entonces podemos simplemente decir que "pseudorrecordamos" las experiencias, o algo por el estilo, y no se perderá nada importante. Pero, en cualquier caso, me parece que una conexión causal con un estado psicológico original relevante es suficiente para que estas creencias puedan considerarse recuerdos.

## 7. El argumento a partir de la referencia*

El tercer argumento en contra de la irrelevancia causal o explicativa de la conciencia es que es incompatible con nuestra capacidad para *referir* a nuestras experiencias conscientes. Parece que podemos pensar acerca de nuestras experiencias conscientes y hablar acerca de ellas... he estado haciendo exactamente esto a través de todo el libro. Pero, a veces, se sostiene que la referencia a una entidad requiere una conexión causal con dicha entidad; esto se conoce como la teoría causal de la referencia. Si esto es así, entonces sería imposible referir a experiencias causalmente irrelevantes.

Sin embargo, no parece haber ninguna razón de principio de por qué la referencia a una entidad *requiere* una conexión causal con esa entidad. Con frecuencia, la referencia involucra una conexión causal, pero no es claro que las cosas deban ser de ese modo. Al hacer referencia a una entidad, todo lo que se requiere es que nuestros conceptos tengan *intensiones* (en particular, intensiones primarias) que la entidad pueda satisfacer. Por ejemplo, mi concepto "la estrella más grande en el universo" tiene una intensión primaria, selecciona un referente en cualquier mundo centrado. En el mundo real, esta intensión selecciona una cierta estrella *S* —esté yo conectado causalmente con *S* o no— de modo que *S* puede considerarse el referente del concepto. Dado que existe una intensión primaria que una entidad en el mundo real debería satisfacer, tenemos los ingredientes básicos necesarios para la referencia.

Ocurre que para muchos de nuestros conceptos, las intensiones primarias se caracterizan causalmente: en un mundo centrado dado,

seleccionan una entidad apropiada que está causalmente conectada con el centro. Esta es la tesis de la denominada teoría causal de la referencia. Pero no hay ninguna razón para que una intensión primaria deba funcionar de este modo. Hay muchas otras funciones que seleccionan, en un mundo centrado hipotético, una entidad que no tiene ninguna conexión causal con el centro. Tales funciones podrían ser intensiones primarias perfectamente adecuadas, con un referente perfectamente apropiado.

Además, la *existencia* de una intensión primaria —aun en casos en los que la intensión primaria se caracteriza causalmente— no depende de ningún modo de una conexión causal con el referente. La intensión primaria es independiente de esos sucesos del mundo real. Una conexión causal puede frecuentemente tener un papel en la *evaluación* de la intensión primaria en un mundo, pero esto es muy diferente de tener un papel en la determinación de la propia intensión primaria. Algunos de nuestros conceptos (por ejemplo, "San Nicolás") no tienen ningún referente en absoluto, pero sí poseen una intensión primaria, una intensión que *podría* haber seleccionado un referente si el mundo hubiese resultado del modo apropiado.

Frecuentemente ocurrirá que la conexión causal con un referente desempeña un papel en la *adquisición* de un concepto, y de esa manera en la formación de una intensión primaria. Se podría argumentar que aun en el caso de "la estrella más grande en el universo", las conexiones causales con el mundo desempeñan un papel en la adquisición de los conceptos básicos a partir de los cuales este concepto composicional se forma. Pero, nuevamente, no parece haber ninguna razón de principio para que la existencia de una intensión primaria *requiera* una conexión causal con el tema relevante. Hasta un cerebro en un tanque podría tener conceptos con intensiones primarias, a pesar de su aislamiento causal (aunque la mayoría de ellas serían intensiones que nada en su mundo satisface). Nuevamente, la *constitución* de una intensión primaria es independiente de conexiones causales de este tipo.

Hay una razón natural de por qué la causalidad es fundamental para tantos de nuestros conceptos: es porque, por lo general, hacemos referencia a aquellas cosas que *conocemos*, y las cosas que conocemos son por lo general cosas con las que estamos causalmente conectados. Pero ya hemos visto que hay buenas razones para rechazar el modelo causal del conocimiento, al menos en el caso de la conciencia: en ese caso tenemos conocimiento de una variedad más inmediata. De este modo, para referir a la conciencia, no necesitamos referir por medio de una intensión que seleccione algo con lo que el centro está causalmente conectado; en cambio, podemos referir por medio de una

intensión que selecciona algo con lo que el centro está inmediatamente familiarizado.

En cualquier caso, lo importante es que 1) mi concepto de "conciencia" puede tener una intensión primaria, ya sea que exista o no una conexión causal con el referente (ya que la existencia de una intensión primaria nunca depende de una conexión causal de este tipo), y 2) la intensión primaria puede seleccionar un referente exista o no una conexión causal con él (ya que no hay ninguna razón para que una intensión primaria deba seleccionar su referente en virtud de una conexión causal). La intensión especifica una función perfectamente apropiada de mundos centrados en características de esos mundos; en este mundo, hay algo que satisface la intensión, de modo que mi concepto tiene un referente. Como hemos visto, la conciencia es algo así como un concepto primitivo (como el espacio y el tiempo, quizá), de modo que no hay ninguna esperanza de caracterizar su intensión en detalle, al modo como podríamos hacerlo con algunos otros conceptos; pero no hay ninguna razón para creer que no debería ser perfectamente capaz de seleccionar un referente en un mundo.

## El contenido de las creencias fenoménicas

Aun si se acepta que el dualismo de propiedades es compatible con la referencia a la conciencia, todavía restan muchos problemas interesantes acerca del contenido de nuestros conceptos y creencias fenoménicas. Primero, tenemos las cuestiones acerca de la naturaleza de las intensiones de nuestros conceptos, tanto para conceptos generales como "conciencia", como para conceptos más específicos como "experiencia de rojo": ¿Exactamente qué seleccionan en un mundo dado? Y segundo, tenemos las cuestiones acerca de qué *constituye* el contenido de nuestros conceptos: ¿Está el contenido constituido solamente por nuestra naturaleza psicológica o por nuestra naturaleza psicológica y fenoménica, y qué papel desempeña cada una? No tengo opiniones definidas sobre estas cuestiones, pero al menos arañaré aquí la superficie.

Un modo interesante de aproximarse a algunas de estas cuestiones es preguntarse si hay alguna diferencia entre el contenido de mis creencias fenoménicas y las de un zombi; y si esto es así, cuál es. Vuelvo aquí a hablar de "creencias" en lugar de "juicios", ya que la pregunta es, precisamente, si existe algún elemento en el contenido de una creencia fenoménica más allá del contenido de un juicio fenoménico, que es, como estipulamos antes, una entidad puramente psicológica. Aceptaré, al menos para el propósito de la discusión, que un zombi tiene creencias (aunque sus creencias no son, ciertamente,

nada más allá de sus juicios). La pregunta es, entonces, si existe alguna diferencia entre el contenido de sus creencias y las mías. En particular, ¿cuál es la diferencia, si existe, entre las condiciones de verdad de nuestras respectivas creencias, y entre las intensiones de nuestros respectivos conceptos?

Un camino que se podría seguir es que el contenido de nuestras creencias y conceptos es exactamente el mismo. Según este enfoque, el zombi tiene conceptos de "conciencia" y de "experiencia de rojo" con las mismas intensiones primarias que mis conceptos correspondientes, y creencias con las mismas condiciones de verdad (primarias). Su concepto "ser consciente", por ejemplo, selecciona a los seres conscientes en un mundo centrado determinado. Sólo que en su mundo no existen seres de ese tipo o, al menos, él no es un ser de esa clase. De modo que su creencia "Soy consciente" tiene las mismas condiciones de verdad (primarias) que mi creencia; la única diferencia es que su creencia es falsa mientras que la mía es verdadera.

Para evaluar este camino, debemos considerar algunas preguntas específicas. Primero, ¿cuándo un zombi dice "Soy consciente", habla falsamente? Algunos dirían que no: debemos interpretar la observación del zombi de un modo caritativo, de modo que su concepto refiera a alguna propiedad funcional que él instancia, y la observación resultará verdadera.[8] Pero el zombi (al menos si es mi gemelo zombi) insistirá que el suyo no es un concepto funcional: él quiere referir a una propiedad suya más allá de su capacidad para discriminar, categorizar, informar, etc. Parece razonable tomarle la palabra en esto. Por supuesto, también podemos permitir una interpretación "deflacionaria" de sus palabras, de modo que aseveraciones como "Recuperé la conciencia" podrían resultar verdaderas en contextos cotidianos; pero, al menos en contextos filosóficos, parece razonable mantener sus conceptos en el alto estándar que él les asigna, de manera, entonces, que sus creencias resultan falsas. No se trata de que padezca una confusión conceptual que podría aclararse mediante un análisis conceptual más cuidadoso. (Si pudiese hacer esto, yo también podría hacerlo; pero esta discusión tiene como premisa la idea de que yo no estoy sufriendo una confusión conceptual de ese tipo.) De modo que parece haber un sentido razonable en el cual sus aseveraciones sobre la conciencia son falsas.[9]

Sin embargo, esto no es suficiente para mostrar que el concepto del zombi tiene la *misma* intensión que el mío. Quizás, esto sólo muestre que tiene un concepto similar al de "propiedad más allá de cualquier propiedad física y funcional", sin que posea el concepto completo de conciencia. La verdadera prueba es si existen mundos centrados en los cuales todo lo relevante es idéntico, pero en los que

alguna creencia del zombi tiene un valor de verdad diferente del mío. Por ejemplo, ¿qué ocurre si ambos estamos hablando con un ser consciente? Yo digo: "Usted es consciente", y hablo verazmente. Si él dice: "Usted es consciente", ¿habla verazmente? Algunos podrían decir que no, porque él carece de la familiaridad inmediata con la conciencia que le permitiría poseer el concepto completo. Pero también existe una intuición de que "sí" podría ser razonable.

Es posible que los ejemplos más relevantes involucren a un ser hipotético con alguna propiedad intrínseca no estructural y no funcional que *no* sea una propiedad fenoménica —si pudiese haber una propiedad así—, y que se encuentra en el tipo usual de relaciones que las propiedades fenoménicas mantienen. Cuando le digo "Usted es consciente" a un ser de este tipo, hablo falsamente; pero, quizá, cuando mi gemelo zombi lo dice, ¿si habló verazmente en el caso anterior, también habla verazmente aquí? En ausencia de cualquier familiaridad con la conciencia, es difícil ver cómo el concepto del zombi podría ser lo suficientemente específico como para distinguir entre los dos casos. Si esto es así, entonces el concepto de conciencia del zombi se queda corto respecto del concepto completo.

Si nuestra familiaridad con la conciencia tiene un papel en la constitución de la intensión primaria de nuestro concepto, esto significaría que el concepto de "conciencia" es diferente de un modo interesante de otros conceptos como "agua". En estos casos la intensión primaria es independiente del referente real; un sistema cognitivo con un referente distinto, o sin ningún referente en absoluto, podría tener la misma intensión primaria. Pero, la conciencia no es un referente ordinario: nuestra relación con ella no está mediada, se encuentra en el centro de nuestra vida mental, de modo que podría tener un papel inusualmente fuerte en la constitución de una intensión primaria. En cualquier caso, dejaré la cuestión abierta.

¿Qué hay de conceptos fenoménicos *específicos*, tales como el concepto de "experiencia de rojo"? Estos son algo más complicados, ya que puede haber más de un concepto en la vecindad de un término de esta clase. Una forma de aproximarse a estos conceptos es considerar cómo describir individuos con el espectro invertido: por ejemplo, alguien que, cuando mira objetos rojos, tiene el tipo de experiencias que yo tengo cuando miro cosas verdes. Cuando digo que una de esas personas (cuando mira una rosa roja) está teniendo una experiencia de rojo, podría existir un sentido relajado en el que la observación pueda considerarse verdadera: en este sentido "experiencia de rojo" se reduce a algo así como "experiencia de la clase típicamente causada (en el individuo que tiene la experiencia) por las cosas rojas". Quizás haya un concepto de "experiencia de rojo" en el lenguaje

público que funciona de un modo similar, pero de cualquier forma lo dejaré de lado.

Más natural, quizá, es un sentido en el cual una observación de este tipo es falsa: la persona no está teniendo una experiencia de rojo sino una experiencia de verde. En una forma de explicitar un sentido de este tipo, la intensión primaria de mi concepto "experiencia de rojo" podría reducirse a algo así como "experiencia de la clase típicamente causada (en mí) por las cosas rojas". Este sentido tiene algunas propiedades interesantes: por ejemplo, si usted y yo tenemos el espectro invertido uno respecto del otro, entonces su concepto "experiencia de verde" seleccionará (lo que yo llamo) experiencias de *rojo*. Se deduce que su observación "El pasto me produce experiencias de verde" podría ser verdadera, aun cuando mi observación "El pasto le produce experiencias de verde" sea falsa. Algunas personas encuentra este resultado desagradable,[10] pero es una consecuencia natural de la indicatividad del concepto (un fenómeno similar surge con las observaciones sobre "mí" y con las observaciones sobre el "agua" hechas por mí y mi gemelo de Tierra Gemela). Si queremos evitar este tipo de cosas, el concepto más "público" mencionado más arriba siempre está disponible.[11]

La intensión primaria de este concepto "experiencia de rojo" debería ser bastante directa, una vez aceptado el concepto general de experiencia. Su concepto "experiencia de rojo" puede tener la misma intensión primaria que el mío, aunque nuestro espectro esté invertido: ambos seleccionan las mismas entidades en un mundo centrado dado (experiencias de la clase típicamente causadas en el centro por cosas rojas), aunque por supuesto nuestros conceptos tendrán *referentes* distintos, ya que habitamos mundos centrados diferentes (el mío está centrado en mí, y el suyo en usted). Debido a esto, nuestros conceptos pueden también tener intensiones secundarias diferentes: los míos seleccionan experiencias de rojo en un mundo contrafáctico, mientras que los suyos seleccionan experiencias de verde. Incluso un zombi podría tener la misma intensión primaria que nosotros dos, al menos si posee el concepto de experiencia, aunque por supuesto su intensión no seleccionará nada en su mundo centrado, y su concepto "experiencia de rojo" no referirá.

Hay más en la historia que esto, sin embargo. Esta interpretación relacional del concepto "experiencia de rojo" es todavía relativamente periférica. Existe un concepto afín —quizás el más importante en estas discusiones —que no se agota en la caracterización relacional. En particular, tenemos un concepto de *cualidad* de las experiencias de rojo. Nada en la caracterización relacional formulada antes captura el concepto de esta cualidad, tal como lo atestigua

el hecho de que la intensión primaria caracterizada anteriormente es compatible con muchas cualidades muy diferentes. Alguien con un espectro invertido compartiría la intensión, pero yo tengo un concepto de esa cualidad —llamémoslo "*R*"— que es *distinto* del concepto correspondiente —llamémoslo "*G*"— de mi contraparte invertida.

A primera vista, podríamos pensar que es suficiente que esa cualidad sea capturada por la intensión secundaria de mi concepto "experiencia de rojo" tal como lo caractericé antes. Como hemos visto, mi contraparte invertida y yo tenemos diferentes intensiones secundarias correspondientes a las diferentes cualidades seleccionadas por nuestros conceptos de "experiencia de rojo". Pero capturarla en una intensión secundaria no es suficiente. Para ver esto, nótese que resulta *informativo* aprender que las experiencias de rojo (esto es, experiencias causadas por objetos rojos) tienen su cualidad específica. O sea, no es en absoluto *a priori* que las experiencias de rojo deban ser *R*. Al aprender esto, restringimos nuestro modelos de cómo es el mundo real: el tipo de experiencia causado por las cosas rojas podría haber sido de *este* modo o de *ese* otro, pero, en cambio, es de *este* modo. Y este tipo de informatividad requiere una diferencia en la intensión primaria: cuando dos conceptos tienen la misma intensión primaria, son *a priori* coextensionales.[12]

Otro modo de ver esto es advertir que cuando María tiene una experiencia de rojo por primera vez, aprende algo *diferente* de lo aprendido por su gemela invertida, que tiene experiencias de verde cuando María las tiene de rojo. María aprende que las cosas rojas causan experiencias como *esta*, mientras que su contraparte aprende que causan experiencias como *esa*. Sus modelos del mundo se estrechan de modos diferentes: María ahora avala un conjunto de mundos centrados, mientras que su gemela avala otro. Se deduce que sus conceptos de las cualidades en cuestión deben tener intensiones primarias diferentes. La intensión primaria de María selecciona experiencias de un tipo (*este* tipo) en cualquier mundo centrado, mientras que su contraparte selecciona experiencias de un tipo diferente.

Mi concepto cualitativo "*R*" no tiene un papel muy directo en las prácticas comunicativas. De ese modo, se parece al "escarabajo en una caja" de Wittgenstein.[13] Mi contraparte invertida tiene un concepto diferente "*G*" relacionado con él, pero otras personas interpretan que dice las mismas cosas que interpretan que yo digo, suponiendo que sus propias situaciones permanecen constantes a través de los dos escenarios. Esto refleja la "inefabilidad" que hice notar en el capítulo 1: a pesar de la rica naturaleza intrínseca de las

sensaciones de rojo, existe poco que pueda decir para comunicar esa diferencia aparte de fijarla por medio de diversas propiedades relacionales, y suponer que los otros tienen las mismas experiencias asociadas. Esto es, parece ser que es el concepto relacional de "experiencia de rojo" el que soporta la carga comunicativa. Podría considerarse que esta inefabilidad proporciona un apoyo indirecto a la irrelevancia explicativa de la experiencia: el hecho de que haya poco que podamos *decir* que capture la cualidad intrínseca de la experiencias concuerda bien con el hecho de que la cualidad no tiene ningún papel directo en la dirección de los procesos cognitivos.[14]

(Por supuesto, todavía podemos *hablar* acerca de esas cualidades, como hice de vez en cuando a través de este libro. Puedo expresar mis creencias fenoménicas en el lenguaje; pero, esto sólo podrá comunicarles a otros el contenido completo de mis creencias si ellos mismos poseen las cualidades relevantes y con las relaciones apropiadas entre ellas.)

Esto claramente proporciona una situación en la cual el contenido de nuestros conceptos y creencias está constituido por algo que va más allá de nuestra estructura física y funcional, de modo que ninguna concepción reductiva del contenido de las creencias tendrá éxito.[15] Mi gemelo invertido y yo podríamos ser físicamente idénticos, pero nuestros conceptos cualitativos correspondientes son distintos, no sólo en la referencia sino también en la intensión primaria. Aquí, aun más claramente que en el caso de la "conciencia", encontramos una situación en la cual el contenido de una creencia fenoménica está constituida por la propia fenomenología. Sucede algo innegablemente interesante: de algún modo, una especie de experiencia, que podríamos pensar como el *referente* de un concepto cualitativo, se introduce en el concepto y constituye su *sentido* (donde sentido se equipara a intensión primaria). Esto es bastante diferente de los casos estándar en los que el objeto de un concepto podría tener un papel en la constitución de una intensión secundaria, pero no de una intensión primaria.[16] Esto es posible sólo por el hecho de que la experiencia está en el *centro* de la mente.

Vemos, entonces, que de hecho hay algo más en una creencia fenoménica que en un juicio fenoménico, al menos en estos casos. Es posible que esto pueda ayudar a comprender la epistemología de la conciencia. Por ejemplo, el caso más específico de este tipo de relación de constitución surgirá cuando una única experiencia $S$ esté involucrada en la constitución de un concepto fenoménico $S$ ("*este* tipo de experiencia").[17] La relación de constitución directa —el modo como la experiencia se introduce en el concepto, por así decirlo— podría ayudarnos a comprender cómo la propia experiencia justifi-

caría una creencia al efecto de que la experiencia es *S*. Esto produce una relación estrecha entre la experiencia y la creencia de una clase que podría pensarse apropiada. Y, dada esta clase de creencia fenoménica justificada específica, es posible ver cómo podrían surgir creencias fenoménicas justificadas más generales (tal como la creencia de que somos conscientes). Dejaré la cuestión aquí, ya que estamos entrando en aguas profundas, pero esta relación entre experiencias y conceptos fenoménicos proporciona mucha materia de reflexión.

(Quizá podríamos formular una tesis al efecto de que una creencia de que una experiencia es *S*, donde el concepto *S* está constituido por la experiencia misma al modo de más arriba, está siempre justificada. Todavía habría creencias fenoménicas injustificadas, pero esto surgiría en casos en los que existe una relación diferente entre el concepto y la experiencia, tal como cuando un concepto constituido por una experiencia —o un conjunto de experiencias, o una descripción extrínseca— se aplica a otra experiencia. Quizás esta tesis relativamente débil pueda capturar un elemento de lo que es plausible en las tesis de "incorregibilidad" estándar, mientras que al mismo tiempo deja espacio para todos los contraejemplos usuales; podría incluso servir como trampolín principal en una concepción detallada de la epistemología de la experiencia. Pero no estoy seguro sobre esto.)

El hecho de que haya un elemento en mi creencia que no está presente en la creencia correspondiente de mi gemelo zombi puede también ayudar a deflacionar cualquier intuición que apoye los argumentos epistemológicos mencionados previamente. *Sólo* son nuestros juicios (funcionalmente interpretados) los que son iguales: no es cierto que las mismas *creencias* habrían estado presentes aun si la experiencia hubiese estado ausente.[18] Y son las creencias, después de todo, las que son fundamentales aquí. Hemos visto que este tipo de diferencia en las creencias no es necesaria para derrotar los argumentos epistemológicos (yo no la supuse en la discusión anterior), pero de todos modos podría ayudarnos a eliminar cualquier duda remanente sobre la cuestión.

Es natural que nos preguntemos cuán lejos podría extenderse este tipo de constitución del contenido por la experiencia. El hecho de que esto es válido para conceptos fenoménicos específicos presta un mayor apoyo a la idea de que también lo es para el concepto más general de "conciencia", aunque la cuestión de la relación entre mi concepto y el de un zombi sigue sin ser clara. Podríamos preguntarnos, entonces, si la experiencia desempeñaría un papel en la constitución del contenido de los conceptos *no* fenoménicos, tales como

conceptos de tipos externos, según algunos filósofos sugirieron. No es obvio cómo debería hacerse la extensión, pero quizás el hecho de que la experiencia desempeñe este papel en un caso pueda darle algún sustento a la idea de que podría hacerlo en otros casos.

De cualquier forma, nada aquí requiere una teoría causal de la referencia. Una conexión causal con la experiencia sería probablemente inapropiada para permitir el tipo de relación directa que encontramos entre las experiencias y las intensiones primarias de los conceptos fenoménicos: en todos los casos usuales en los que existe una conexión causal ("agua", por ejemplo), no existe una relación de este tipo. En cambio, parece ser nuestra familiaridad inmediata con la experiencia la que hace que este tipo de constitución sea posible. De modo que no se requiere una conexión causal con la experiencia para constituir la posesión de las intensiones primarias relevantes; y ciertamente no se requiere ninguna conexión causal para que las intensiones primarias seleccionen un referente en estos casos; de manera que no existe ninguna dificultad de principio en referir a la experiencia, aun en un enfoque dualista de propiedades.

Parece, entonces, que a pesar de que la irrelevancia explicativa de la experiencia para la conducta física puede, al principio, ser contraria a la intuición, no existe ningún argumento fuerte en su contra. Los que podrían parecer argumentos fuertes, los realizados a partir de la epistemología y de la referencia, resultan ser meros desafíos. La consideración de estos puntos plantea un gran número de cuestiones interesantes, pero finalmente hemos visto que hay buenas razones para creer que la epistemología y la semántica de la experiencia no pueden ser esencialmente causales y deberían, en cambio, entenderse en otros términos. He dicho poco aquí acerca de cómo podríamos comprender estas cosas en un enfoque dualista de propiedades. Una comprensión completa de estas cuestiones requeriría una extensa investigación separada; pero espero haber dicho lo suficiente como para dejar en claro que el enfoque no reductivo proporciona un marco natural para encontrarle un sentido a estas cuestiones.

# PARTE III

# HACIA UNA TEORIA DE LA CONCIENCIA

# 6

# La coherencia entre la conciencia y la cognición

## 1. Hacia una teoría no reductiva

Aunque no es posible explicar reductivamente la conciencia, puede existir una teoría de ella. Simplemente debemos cambiar por una teoría *no reductiva*. Podemos desistir del proyecto de intentar explicar totalmente la existencia de la conciencia en términos de algo más básico y, en cambio, admitir que es fundamental y formular una concepción de cómo se relaciona con todo lo demás en el mundo.

Una teoría de esta clase será de un tipo similar al de las teorías físicas de la materia, el movimiento o el espacio y el tiempo. Las teorías físicas no derivan la existencia de estas características de ninguna otra cosa más básica, sino que formulan concepciones detalladas y sustanciales de ellas y de cómo se interrelacionan; esto da como resultado explicaciones satisfactorias de muchos fenómenos específicos que involucran a la masa, el espacio y el tiempo. Lo hacen mediante la formulación de un conjunto simple y potente de *leyes* que emplean esas diversas características, de las que todo tipo de fenómenos específicos surgen como consecuencia.

Por analogía, la piedra angular de una teoría de la conciencia será un conjunto de *leyes psicofísicas* que gobiernen la relación entre la conciencia y los sistemas físicos. Ya hemos aceptado que la conciencia superviene naturalmente (aunque no lógicamente) a lo físico. Esta superveniencia debe estar asegurada por las leyes psicofísicas; una concepción de esas leyes nos dirá exactamente *cómo* la conciencia depende de los procesos físicos. Dados los hechos físicos acerca de un sistema, esas leyes nos permitirán inferir qué tipo de experiencia consciente estará asociada con el sistema, si es que alguna lo está.

Estas leyes estarán al mismo nivel que las leyes de la física como parte del mobiliario básico del universo.

Se deduce que, aunque esta teoría no explicará la existencia de la conciencia en el sentido de decirnos "por qué existe la conciencia", podrá explicar instancias específicas de ella en términos de la estructura física subyacente y las leyes psicofísicas. Nuevamente, esto es análogo a la explicación en la física, que da cuenta de por qué instancias específicas de la materia o el movimiento tienen las características que tienen, invocando principios subyacentes generales en combinación con ciertas propiedades locales. Todo tipo de fenómenos físicos macroscópicos pueden explicarse en términos de las leyes físicas subyacentes; de modo similar podríamos esperar que todo tipo de fenómenos experienciales "macroscópicos" puedan ser explicados por las leyes psicofísicas de una teoría de la conciencia.

No es necesario que haya nada sobrenatural acerca de estas leyes. Son parte del mobiliario básico de la naturaleza, como lo son las leyes de la física. Habrá algo "primitivo" en ellas, es cierto. En algún nivel, las leyes deberán ser consideradas verdaderas y no explicadas más allá. Pero lo mismo ocurre en la física: las leyes últimas de la naturaleza siempre, en algún punto, parecerán arbitrarias. Es esto lo que las hace leyes de la naturaleza y no leyes de la lógica.

En la ciencia, nunca obtenemos algo por nada: alguna cosa, en algún lugar debe presuponerse. Es un hecho notable que en la mayor parte de las áreas de la ciencia, todo lo que en última instancia debemos presuponer son las leyes de la física y, quizás, algunas condiciones fronterizas. Pero no hay ninguna razón de que las leyes de la física deban ser absolutamente privilegiadas de ese modo. Si resulta que en el estudio de la conciencia debemos presuponer algún aspecto de la relación entre los procesos físicos y la conciencia, entonces que así sea. Ese es el precio de construir una teoría.

No obstante, siempre queremos presuponer tan poco como podamos. Una teoría fundamental no dejará la conexión en el nivel de "El estado cerebral X produce el estado consciente Y" para una vasta colección de estados físicos complejos y experiencias asociadas. En cambio, sistematizará esa conexión por medio de un marco explicativo subyacente, especificando leyes simples en virtud de las cuales es válida la conexión. La física no se contenta con ser una mera masa de observaciones acerca de las posiciones, velocidades y cargas de diversos objetos en diversos momentos; sistematiza estas observaciones y muestra cómo resultan ser consecuencias de leyes subyacentes, que son tan simples y potentes como sea posible. Lo mismo debería ser válido para una teoría de la conciencia. Deberíamos intentar explicar

la superveniencia de la conciencia a lo físico en términos del conjunto más simple posible de leyes.

Finalmente, querremos un conjunto de *leyes fundamentales.* Los físicos buscan un conjunto de leyes básicas lo suficientemente simples como para poder escribirlas en el frente de una remera; en una teoría de la conciencia, deberíamos esperar lo mismo. En ambos casos, buscamos la estructura básica del universo y tenemos buenas razones para pensar que esta estructura es de una notable simpleza. Sin embargo, el descubrimiento de leyes fundamentales puede ser una meta distante. En la física, primero tuvimos leyes que caracterizaban las regularidades macroscópicas y solamente después accedimos a las leyes fundamentales subyacentes. En una teoría de la conciencia, podemos esperar que ocurra lo mismo: partiríamos de leyes *no básicas* que caracterizan la relación entre los procesos físicos y la experiencia consciente en un nivel relativamente alto. Esta clase de principios de alto nivel podría ofrecernos un apoyo explicativo significativo en el ínterin, así como se utilizaron los principios de la termodinámica mucho antes de que hubiésemos descubierto los principios subyacentes de la mecánica estadística. Estas leyes de alto nivel, una vez descubiertas, plantearán fuertes restricciones sobre cualquier ley fundamental subyacente, guiándonos así en la búsqueda de una teoría última.

Cuando finalmente tengamos acceso a teorías fundamentales de la física y de la conciencia, podremos tener lo que verdaderamente sería una teoría de todo. Las leyes físicas fundamentales explicarán las características de los procesos físicos; las leyes psicofísicas explicarán las experiencias conscientes asociadas; y todo el resto será una consecuencia.

Por supuesto, podría ocurrir que en la búsqueda de estas teorías, ocurran desarrollos que cambien nuestra concepción de lo que sería la teoría última. Podría ocurrir, por ejemplo, que encontremos leyes generales que subsumen los fenómenos de la física y de la conciencia en una teoría más amplia, así como encontramos una teoría que subsumió la electricidad y el magnetismo, y así cómo los físicos buscan en la actualidad una teoría que unifique todas las fuerzas físicas fundamentales. Quizás ocurran desarrollos que sean todavía más sorprendentes. Pero el marco conceptual vigente hoy nos proporciona al menos un punto de partida en la búsqueda de una teoría que serviría como primera aproximación y trampolín para cualquier teoría sucesora más radical que pudiese surgir en el futuro.

## ¿Cómo podríamos construir una teoría de la conciencia?

Toda esta grandiosidad metafísica está bien, podríamos responder, ¿pero cómo traducirla en la práctica? En particular, ¿cómo podemos *descubrir* las leyes psicofísicas que constituirán una teoría de la conciencia? Después de todo, existe un enorme problema para una teoría de la conciencia que una teoría de la física no enfrenta: la ausencia de datos. Debido a que la conciencia no es directamente observable en contextos experimentales, no podemos simplemente realizar experimentos que midan las experiencias asociadas a los diversos procesos físicos y así confirmar o disconfirmar diversas hipótesis psicofísicas. ¿Es posible que estas leyes, aun si existiesen, pudieran permanecer en un limbo incognoscible? Podría pensarse que la no verificabilidad de cualquier teoría de la conciencia que podamos formular relegaría estas teorías al estado de una pseudociencia.

Ciertamente, esta preocupación es legítima: es esto lo que hace que una teoría de la conciencia sea más difícil de captar que una de la física. Pero ello no nos impide en absoluto buscar una teoría de la conciencia. Para empezar, cada uno de nosotros tiene acceso a una rica fuente de datos en nuestra propia persona. Conocemos nuestras experiencias conscientes de un modo detallado y específico, y también conocemos los procesos físicos subyacentes, de modo que existe un conjunto significativo de regularidades precisamente allí. Dadas esas regularidades, podemos invocar algún tipo de inferencia de la mejor explicación para encontrar las leyes subyacentes más simples posibles que pudiesen generarlas. En este momento, no poseemos un conjunto de leyes capaces de hacer esto, de modo que esta no es en absoluto una restricción trivial sobre una teoría. Bien podría resultar que sólo exista un conjunto razonablemente simple de leyes que produzca los resultados correctos, en cuyo caso tendríamos buenas razones para creer que esas leyes son parte de una teoría correcta.

No obstante, podría objetarse que teorías de todo tipo pueden mantener compatibilidad con los datos de primera persona: desde teorías solipsistas (en las que sólo yo soy consciente) a teorías panpsíquicas (en las cuales todo es consciente); desde teorías bioquímicas (en las cuales la conciencia surge sólo a partir de ciertas organizaciones bioquímicas) a teorías computacionales (en las cuales la conciencia surge de cualquier ente con el tipo correcto de organización computacional); incluyendo también teorías extravagantes como la teoría de que las personas sólo son conscientes en los años impares (escribo esto en 1995). ¿Cómo podemos desechar cualquiera de esas teorías, si no podemos asomarnos dentro de la mente de otros para medir su experiencia consciente?

Todas esas teorías son *lógicamente compatibles* con los datos, pero esto no basta para hacerlas *plausibles*. Las teorías solipsistas, por ejemplo, son extremadamente inverosímiles debido a su gran arbitrariedad (¿por qué sólo *esta* persona debería ser consciente?) y su gran heterogeneidad en el espacio y el tiempo (mi experiencia consciente está sistemáticamente vinculada a mi estructura física, pero un duplicado físico mío localizado en otro lugar no será consciente en absoluto). En toda clase de dominios, consideraciones de plausibilidad de todo tipo tienen un papel en la conformación de nuestras teorías, más allá del papel desempeñado por la evidencia empírica. Considérese, por ejemplo, nuestra aceptación de la teoría de la evolución, a diferencia de la teoría de que el mundo fue creado hace cincuenta años con recuerdos y registros fósiles intactos. O considérese la aceptación de ciertas teorías simples de la mecánica cuántica en lugar de otras empíricamente equivalentes pero altamente "retocadas". La evidencia empírica no es todo lo que tenemos para proceder a la formación de teorías; también existen principios de plausibilidad, simplicidad y estética, entre otras consideraciones.

El papel desempeñado por la simplicidad, en particular, no puede ser exagerado. Sin esta restricción, la teorización científica en general estaría tristemente subconstreñida. Para cualquier teoría científica se puede fácilmente construir una hipótesis *ad hoc* que sea empíricamente equivalente. Sin embargo, nadie aceptará una hipótesis de este tipo precisamente debido a su innecesaria complejidad. De modo que si podemos encontrar un conjunto simple de leyes subyacentes que sean compatibles con los datos disponibles, tenemos buenas razones para rechazar las alternativas más complejas.

Otras restricciones de plausibilidad podrían tener una gran importancia para la generación de una teoría de la conciencia. La más obvia es el principio en el que nos basamos cada vez que tomamos el informe verbal de alguien como un indicador de su experiencia consciente: los informes de esas personas acerca de sus experiencias en general reflejan precisamente su contenido. Este no es un principio que podamos probar que sea verdadero, pero es *a priori* mucho más plausible que la alternativa. Esta plausibilidad se basa, en cierta medida, en una inferencia a partir de nuestro propio caso, pero también tiene el carácter de una restricción metodológica sobre el desarrollo de una teoría de la conciencia. Si el principio resultase ser totalmente falso, todas las apuestas serían inútiles: en ese caso, el mundo simplemente sería un lugar irrazonable y una teoría de la conciencia estaría más allá de nuestras posibilidades. Cuando desarrollamos cualquier clase de teoría, suponemos que el mundo es un lugar razonable, en el que los planetas no cobran existencia repen-

tinamente con registros fósiles completamente formados y donde las leyes complejas no están "arregladas" para reproducir las predicciones de otras más simples. De otro modo, cualquier cosa valdría.

Gracias a un supuesto de plausibilidad como este, contamos con una restricción muy útil sobre una teoría de la conciencia y ciertamente con una fuente muy rica de datos, incluso del caso de tercera persona. Para descubrir si alguien está experimentando conscientemente un estímulo, ¡sencillamente pregúntele! Este principio nos permite extraer conclusiones mucho más fuertes acerca de la asociación entre las experiencias conscientes y sus bases físicas. Desde luego, el supuesto es tan plausible que los investigadores se apoyan en él todo el tiempo, y pocos pensarían en cuestionarlo. Supuestos relacionados pueden también desempeñar un papel valioso, como el principio de que los recuerdos que las personas poseen de sus experiencias no son por lo general radicalmente incorrectos. Por supuesto, pueden existir excepciones ocasionales a estos principios; pero, al menos, existe la presunción de que es probable que los informes y recuerdos sean reflexiones precisas de la experiencia en ausencia de buenas razones para creer otra cosa.

Otros supuestos de plausibilidad podrían incluir los siguientes: que las leyes fundamentales son homogéneas en el espacio y el tiempo; que la experiencia consciente sólo depende del estado físico interno de un organismo; que es improbable que factores arbitrarios como la distribución de las moléculas en las neuronas se reflejen en la experiencia consciente, a menos, quizá, que afecten su funcionamiento, etc. Desde luego, es lógicamente posible que cualquiera de estos supuestos sea falso pero, en ausencia de razones para creer otra cosa, es razonable suponer que son verdaderos. Juntos, estos supuestos de plausibilidad plantean fuertes restricciones sobre una teoría de la conciencia y pueden ayudarnos considerablemente a generar una teoría de este tipo.

¿Qué hacer entonces con la preocupación de que una teoría de la conciencia no es verificable? Esta preocupación sólo tendrá un efecto importante si resulta que existen dos teorías igualmente simples, las dos son perfectamente compatibles con los datos, y ambas satisfacen las restricciones de plausibilidad pertinentes. Esto bien podría no ocurrir: podría surgir una única teoría que sea claramente superior a todas las competidoras. Si surgen dos teorías igualmente buenas, podríamos estar en dificultades para elegir entre ellas, pero aun así habremos adquirido una comprensión significativa de la conciencia al constreñir las cosas hasta ese punto. En cualquier caso, es claramente prematuro preocuparse por la no verificabilidad antes de tener al

menos una teoría que pueda tratar los fenómenos de un modo aunque más no sea remotamente satisfactorio.

Por supuesto, el hecho de depender de datos de primera persona y de restricciones de plausibilidad significa que una teoría de la conciencia tendrá un carácter especulativo no compartido por las teorías de la mayoría de los dominios científicos. Debido a que la verificación intersubjetiva rigurosa es imposible, nunca nos encontraremos tan seguros de que nuestras teorías están en el camino correcto. Por esta razón, es probable que la ciencia de la conciencia nunca poseerá las firmes credenciales empíricas de otras ciencias, y que la mayoría de los investigadores obstinados siempre se mantendrán a distancia. Sin embargo, la conciencia es un fenómeno tan fundamental que es mejor tener alguna comprensión de ella que ninguna. Si puede diseñarse una teoría razonable de la conciencia y resulta ser superior a todas sus competidoras, esto será un logro de cierta importancia, aun cuando la teoría pueda no llegar nunca a recibir un apoyo del todo concluyente. Simplemente, este es el barco en el que nos encontramos cuando tratamos de comprender el universo: tomamos los materiales que tenemos y trabajamos con ellos.

En este capítulo y los próximos dos, daré algunos pasos iniciales hacia una teoría de la conciencia. Los primeros dos capítulos analizan posibles leyes psicofísicas *no básicas*; argumentaremos en favor de ciertos principios que expresan regularidades de alto nivel en la dependencia de la conciencia sobre los procesos físicos. En el tercer capítulo especularemos acerca de la naturaleza de las leyes fundamentales subyacentes. Todo esto tiene el carácter de una investigación preliminar, pero debemos empezar por algún sitio.

## 2. Principios de coherencia

El modo más prometedor de comenzar a desarrollar una teoría de la conciencia es concentrarse en la notable *coherencia* entre la experiencia consciente y la estructura cognitiva. La fenomenología y la psicología de la mente no son independientes una de la otra; están relacionadas de un modo sistemático. Las muchas relaciones legaliformes entre la conciencia y la cognición pueden proporcionar mucho de lo que necesitamos para lograr que una teoría de la conciencia despegue. El mejor modo de aprehender esta relación es concentrarnos en los juicios fenoménicos. Estos juicios son parte de la psicología, pero están estrechamente ligados a la fenomenología y, como tales, representan un puente entre los dominios. Si pensamos acerca de estos juicios y del modo como funcionan en nosotros,

podremos obtener un número de principios que conectan lo fenoménico con lo psicológico.

El principio más obvio de este tipo es el que mencioné en el apartado 1: nuestros juicios de segundo orden sobre la conciencia son, por lo general, correctos. Podemos llamar a esto el principio de *fiabilidad.* Cuando juzgo que estoy experimentando una sensación auditiva, usualmente estoy experimentando una sensación auditiva. Cuando creo que acabo de experimentar un dolor, usualmente acabo de experimentarlo. Existe también un principio recíproco, que llamaré el principio de la *detectabilidad*: cuando ocurre una experiencia, por lo general tenemos la capacidad de formar un juicio de segundo orden acerca de ella. Por supuesto, muchas experiencias ocurren sin que les prestemos ninguna atención, pero usualmente tenemos la *capacidad* de advertirlas: una experiencia que en principio fuera imperceptible para nosotros sería de una clase extraña.[1]

Los principios que esbocé no son absolutos. Nuestros juicios de segundo orden pueden a veces ser erróneos, lo que provee excepciones al principio de fiabilidad. Esto podría ocurrir debido a fallos en la atención (si estoy distraído, puedo creer que acabo de experimentar un dolor cuando en realidad sólo experimenté un ruido fuerte), dificultades en la captación de las categorías pertinentes (como cuando rotulo equivocadamente una experiencia de carmesí como de castaño), enfermedad mental o patología neurofisiológica (como en los casos de negación de ceguera, en la que los sujetos hacen afirmaciones falsas acerca de sus experiencias), y por otras diversas razones. En la dirección opuesta, podría sostenerse que las experiencias pasarían inadvertidas si ocurren mientras estamos dormidos, por ejemplo, o si destellan demasiado rápido como para que alcancemos a percibirlas. De todas formas, estos principios al menos encierran regularidades significativas. En un caso típico, un juicio de segundo orden por lo general será correcto y una experiencia usualmente será discernible. Estas regularidades no son leyes sin excepciones, pero son válidas con demasiada frecuencia como para que se trate de una mera coincidencia. Algo sistemático ocurre.

No intentaré justificar estas afirmaciones en detalle, ya que no constituyen los principios de coherencia fundamentales de los que me ocuparé. Pero la consideración del caso de primera persona hace evidente que los principios son plausibles al menos allí, y pueden extenderse naturalmente a otros casos mediante los principios de homogeneidad y simplicidad. Los principios también son respaldados por el sentido común, el que posee un cierto peso; por supuesto, es posible prescindir del sentido común si tenemos buenas razones para hacerlo, pero si todo el resto es igual, deberíamos colocarnos del lado

del sentido común en lugar de en su contra. Finalmente, como hice notar antes, estos principios tienen la naturaleza de una especie de restricción metodológica sobre el desarrollo de una teoría de la conciencia. Si los juicios de segundo orden no fuesen fiables en general o si la mayoría de nuestras experiencias fueran totalmente imperceptibles, entonces nuestros juicios acerca de la experiencia tendrían tan poca relación con la realidad que una teoría de la conciencia nunca podría despegar.

## La coherencia entre la conciencia y la percatación

El principio de coherencia más importante entre la conciencia y la cognición no involucra juicios fenoménicos de segundo orden. Más bien, concierne a la relación entre la conciencia y los juicios de primer orden. Los principios de los que nos ocupamos aquí conciernen a la coherencia entre la conciencia y la *percatación*. Recuérdese que la percatación es el correlato psicológico de la conciencia, algo que puede explicarse en forma aproximada como un estado en el que alguna información es directamente accesible y está disponible para el control deliberado de la conducta y para su información verbal. El contenido de la percatación corresponde al contenido de los juicios fenoménicos de primer orden (con una salvedad que mencionaremos más adelante), es decir, a los estados con contenido—como "Ese objeto es rojo"— que no son acerca de la conciencia, sino paralelos a ella.

Allí donde hay conciencia, hay percatación. Mi experiencia visual de un libro rojo sobre mi escritorio está acompañada por una *percepción* funcional del libro. La estimulación óptica se procesa y transforma, y mis sistemas perceptuales registran que existe un objeto de tal y tal forma y color sobre el escritorio, y esta información está disponible para el control de la conducta. Lo mismo ocurre para los detalles específicos de lo que se experimenta. Cada detalle está cognitivamente representado en la percatación. Para ver que cada detalle debe estar así representado, simplemente obsérvese que yo puedo hacer comentarios sobre esos detalles y dirigir mi conducta en modos que dependen de ellos; por ejemplo, puedo señalar partes apropiadas del libro. Esta disponibilidad sistemática de la información implica la existencia de un estado interno que posee ese contenido.

Este estado interno es un juicio fenoménico de primer orden, al menos en una primera aproximación. Incluyo esta salvedad porque se podría cuestionar si ese estado debería estrictamente denominarse un "juicio". Su contenido no tiene por qué ser algo que el sujeto confirmaría si reflexionase sobre ello e incluso podría no conceptuali-

zarlo en absoluto. Un estado de esta clase sólo podría considerarse un juicio en un sentido débil; podría ser mejor hablar de él como una especie de registro informacional o, en el mejor de los casos, como un juicio implícito o subpersonal. Más adelante en el capítulo analizaré esta cuestión con mayor detalle pero, por ahora, cuando hable de estos estados como juicios, debe entenderse la expresión en un sentido amplio: selecciona una clase de estados representacionales que no necesitan ser confirmados reflexivamente por el sujeto y que no necesitan tampoco poseer un contenido conceptualizado.

Lo que es válido aquí para la experiencia visual lo es también para cualquier experiencia sensorial. Lo que experimentamos en la audición se representa en nuestro sistema auditivo de una manera tal que los procesos subsiguientes tienen acceso a esas representaciones para el control de la conducta; en particular, el contenido está disponible para ser comunicado verbalmente. En principio, alguien que no supiese nada acerca de la conciencia podría examinar nuestros procesos cognitivos y determinar los contenidos de la percatación observando el papel que esa información tiene en la dirección de los procesos subsiguientes. Podemos tratar del mismo modo las alucinaciones y otros casos de sensaciones en los que no se percibe un objeto real. Aunque no existe ningún objeto real del que pueda ocuparse la percepción, todavía encontramos una representación en nuestro sistema perceptual. Macbeth tenía un estado cognitivo de primer orden con el contenido "daga allí" que acompañaba su experiencia de una daga, a pesar del hecho de que no había ninguna daga para ser percibida o experimentada.

La experiencia no perceptual también pertenece a esta misma categoría. A pesar de que puede no haber ningún *objeto* en una experiencia de dolor, los contenidos del tipo "algo duele" —o aun mejor, "algo malo"— también tienen una representación cognitiva. El que podamos hacer comentarios acerca del dolor y que podamos dirigir nuestra conducta apropiadamente pone de relieve este hecho. Hay percatación aquí así como hay percatación en la percepción visual, aun cuando el objeto de esta percatación no esté tan bien definido. Algo similar ocurre con nuestra experiencia de las emociones y otras experiencias "internas". En todos estos casos, existen estados cognitivos que corresponden a las experiencias; si no los hubiese, entonces el contenido de la experiencia no podría en absoluto reflejarse en la conducta.

Obsérvese que el principio no es que cada vez que tenemos una experiencia consciente nos percatamos *de esa experiencia*. Son los juicios de primer orden los que son fundamentales aquí, no los juicios de segundo orden. El principio es que cuando tenemos una experien-

cia, nos percatamos del *contenido* de la misma. Cuando experimentamos un libro, nos percatamos del libro; cuando experimentamos un dolor, nos percatamos de algo doloroso; cuando experimentamos un pensamiento, nos percatamos de lo que sea que el pensamiento trate. No es una cuestión de una experiencia seguida de un juicio separado, como podría ser el caso para los juicios de segundo orden; los juicios de primer orden son concomitantes de las experiencias, existen junto con ellas.

El vínculo entre las experiencias y los juicios de segundo orden es mucho más indirecto: aunque tenemos la *capacidad* de advertir nuestras experiencias, la mayor parte del tiempo sólo notamos su contenido, no las propias experiencias. Es sólo ocasionalmente que nos relajamos y notamos nuestra *experiencia* del libro rojo; usualmente sólo pensamos en el libro. Mientras que los juicios de segundo orden son poco frecuentes, los juicios de primer orden ocurren permanentemente. El vínculo más directo es, por lo tanto, el vínculo entre la conciencia y los juicios de primer orden.

Hasta ahora argumenté que allí donde hay conciencia hay percatación. Pero la flecha va en ambos sentidos. Donde hay percatación, en general hay conciencia. Cuando estamos conscientes de algo en nuestro ambiente, cuando algún contenido informable dirige nuestra conducta, existe por lo general una experiencia consciente correspondiente. Cuando mi sistema cognitivo representa un perro ladrando, tengo una experiencia de un perro ladrando. Cuando me percato del calor en torno mío, siento calor, y así sucesivamente.

Las cosas se ponen un poco difíciles aquí. Podría parecer que existen diversos tipos de percatación que no tienen experiencias correspondientes. La percatación que involucra la información en la memoria proporciona un ejemplo. Yo me percato de que Clinton es presidente, en el sentido de que tengo acceso a esa información, puedo comunicarla verbalmente y puedo utilizarla en la dirección deliberada de mi conducta. Sin embargo, si no tengo un pensamiento ocurrente al efecto, no parece que pueda existir una experiencia consciente correspondiente; o, si la hay, es una experiencia extraordinariamente débil. De un modo similar, yo podría percatarme (no ocurrentemente) de que hay una bicicleta afuera, sin que exista una experiencia asociada de bicicleta. Este tipo de percatación sin experiencia es más pronunciada con la percatación proposicional —yo me percato *de que* mi bicicleta está afuera— aunque también se aplica a un tipo de percatación objetal, en el sentido de que parece razonable decir que me percato *de mi bicicleta.*

Podríamos dejar la cuestión tal como está, pero es más satisfactorio formular restricciones sobre la noción de percatación para

mejorar su paralelismo con la conciencia. Resulta plausible que exista *algún* tipo de diferencia funcional entre los procesos involucrados en un tipo de caso y en el otro: el propio hecho de que puedo informar acerca de la diferencia entre ellos es un testimonio de esto. Es esta diferencia funcional la que debe ser aislada.

Quizás, la diferencia más importante sea que en los casos de percatación con conciencia existe un tipo de acceso *directo* del que carecen los casos de percatación sin conciencia. Por ejemplo, debo "evocar" la información de que Clinton es presidente para que marque una diferencia en el control deliberado de la conducta, por lo menos si no es el contenido de un pensamiento ocurrente. No está inmediatamente preparado para marcar una diferencia en el control, en contraposición a los estados cognitivos asociados con experiencias y pensamientos ocurrentes. Es decir, el acceso cognitivo a la información es, en este caso, algo más indirecto. Es esto lo que proporciona la distinción funcional entre los pensamientos ocurrentes y no ocurrentes.

Podemos, entonces, incorporar esta inmediatez de acceso en una noción revisada de la percatación. De acuerdo con la noción revisada, los pensamientos no ocurrentes no califican como parte del contenido de la conciencia, pero los pensamientos ocurrentes sí. Respectivamente, deberíamos esperar que los pensamientos ocurrentes estén asociados con experiencias, aun cuando los pensamientos no ocurrentes no lo estén. Esto es precisamente lo que encontramos. Mi pensamiento no ocurrente de que Clinton es presidente no tiene ningún impacto sobre mi fenomenología, pero un pensamiento ocurrente al efecto estará asociado con una experiencia. Para ver esto, nótese que hay *algo* que es como pensar para uno mismo que Clinton es presidente; si yo no hubiese estado pensando ese pensamiento justo ahora, habría sido algo sutilmente diferente ser yo mismo.[2]

De esta forma, es plausible que, si definimos la percatación de un modo apropiado, la conciencia esté siempre acompañada de la percatación y viceversa. Hay más para decir sobre la caracterización del tipo apropiado de percatación; en lo que sigue refinaré aun más esta caracterización, en parte mediante la consideración de diversos casos interesantes. Aun en un nivel aproximado, sin embargo, podemos ver que esta relación proporciona un punto focal útil para la comprensión de la coherencia entre la conciencia y la cognición.

## El principio de coherencia estructural

Hasta ahora tenemos una hipótesis: allí donde hay conciencia, hay percatación y donde hay (el tipo correcto de) percatación, hay conciencia. La correlación entre estos dos elementos puede detallarse

aun más. En particular, diversas características *estructurales* de la conciencia corresponden directamente a características estructurales que están representadas en la percatación.

La experiencia consciente de un individuo no es en general una masa informe homogénea; posee una estructura interna detallada. Mi campo visual, por ejemplo, tiene una geometría definida implícita. Existe un fragmento rojo grande aquí, con un fragmento amarillo pequeño en la proximidad, con algo de blanco en medio; existen patrones de franjas, cuadrados y triángulos, etc. En tres dimensiones, tengo experiencias de formas como cubos, experiencias de qué cosa está detrás de cuál otra y otras manifestaciones de la geometría de la profundidad. Mi campo visual consiste en una vasta masa de detalles que encajan en una estructura general.

Crucialmente, todos estos detalles están cognitivamente representados dentro de lo que podemos concebir como la estructura de la percatación. Por ejemplo, el tamaño y la forma de los diversos fragmentos están representados en mi sistema visual, quizás en un mapa topográfico relativamente directo, pero aunque no sea así, sabemos que están representados de algún modo. Deben estarlo, como lo atestigua el hecho de que la información pertinente está disponible para guiar el control de la conducta. Lo mismo ocurre para la representación perceptual de las franjas y las formas cúbicas, etc. Cada uno de estos detalles estructurales es accesible al sistema cognitivo y está disponible para su utilización en el control de la conducta, de modo que cada uno está representado en el contenido de la percatación.

En principio, alguien que posea un conocimiento completo de mis procesos cognitivos podría recuperar todos esos detalles estructurales. La geometría del campo visual puede recuperarse mediante un análisis de la información que el sistema visual pone a disposición de los procesos de control posteriores; el hecho de que cada uno de estos detalles puede reflejarse en las capacidades conductuales del sujeto —un sujeto podría trazar los diversos detalles estructurales con movimientos de brazos, por ejemplo, o comentar sobre ellos en informes verbales— implica que la información debe estar presente en algún lado. Por supuesto es probable que los detalles del análisis sean bastante complicados y se encuentren mucho más allá de las posibilidades de los métodos actuales, pero sabemos que la información está allí. De este modo, podemos ver que la estructura de la conciencia se refleja en la estructura de la percatación.

Lo mismo ocurre para la estructura *implícita* en el campo fenoménico, tal como las relaciones entre los colores. Aunque sólo esté viendo un color en un momento dado, existe una multitud de colores

que *podría* estar viendo, colores con los que este mantiene una relación estructural. Un color es muy similar a otro, y bastante diferente de otro. Dos colores pueden parecer complementarios, o un grupo puede parecer "cálido" y otro grupo "frío". En un análisis detallado, nuestros colores fenoménicos se situarían en una estructura tridimensional organizada según una dimensión rojo-verde, una dimensión amarillo-azul y una dimensión blanco-negro (la elección de los ejes es algo arbitraria, pero siempre habrá tres de ellos). Resulta que esta estructura fenoménica tridimensional se refleja en una estructura tridimensional de la información del color que procesan nuestros sistemas perceptuales. Por supuesto, es predecible que así ocurra, ya que sabemos que la información pertinente está disponible para el control de la conducta, pero resulta interesante ver que existen estudios del sistema visual (véase Hardin 1988 para un análisis) que actualmente se encuentran elaborando en detalle esa estructura. Podríamos decir, en este caso, que existe una *estructura diferencial* en nuestra experiencia consciente (un espacio de diferencias entre las experiencias posibles) que se refleja en una estructura diferencial en la percatación: a la multiplicidad de experiencias de color y relaciones entre ellas le corresponde respectivamente una multiplicidad de representaciones de color y relaciones entre ellas.

Podemos encontrar tipos similares de estructuras implícitas en otros dominios fenoménicos y una correspondencia similar con las estructuras implícitas en el nivel de procesamiento. La estructura fenomenológica en un acorde musical debe reflejarse en la estructura de lo que está representado, por ejemplo, para que se pueda informar sobre ella o se refleje en otros procesos de control. Lo mismo ocurre con la estructura implícita de los sabores. Estas correspondencias se encuentran en los estudios empíricos de los procesos pertinentes con mucha frecuencia; pero aun sin esos estudios, podemos notar que debe existir algún tipo de correspondencia si reflexionamos sobre el hecho de que esos detalles estructurales están disponibles para desempeñar un papel de control. En general, este tipo de razonamiento nos lleva a la conclusión de que cualquier estructura detallada que podamos encontrar en un campo fenoménico se reflejará en las estructuras representadas en la percatación.

Existen varias características más específicas de la experiencia que también se reflejan en la percatación. La más obvia de estas es la *intensidad* de la experiencia. Es evidente que la intensidad marca una diferencia en el procesamiento posterior, de manera que debe estar representada de alguna forma en la estructura de la percatación. Es plausible que la intensidad de una experiencia tenga una correspondencia directa con el grado en el cual una representación subya-

cente tiende a desempeñar un papel de control y ocupar los recursos de procesamiento posteriores (piénsese en la diferencia entre un dolor intenso y uno débil, o entre una emoción devastadora y una tono emotivo). Otra característica de esta clase es la *resolución* de las experiencias, tal como puede encontrarse, por ejemplo, en la diferencia entre la alta resolución en el centro de un campo visual y la baja resolución en la periferia. Esta resolución es algo que deberíamos esperar que se refleje en la resolución de las representaciones subyacentes y, precisamente, eso es lo que encontramos.

En general, aunque las experiencias son, en algún sentido, "inefables", las relaciones entre las experiencias no lo son; no tenemos ningún problema para analizar estas relaciones, sean de similitud y diferencia, geométricas, de intensidad, etc. Como Schlick (1938) señaló, la *forma* de la experiencia parece ser directamente comunicable, aun cuando el *contenido* (cualidad intrínseca) no lo sea: puedo caracterizar la relación entre una experiencia de rojo y una de verde, aunque no pueda hacer lo mismo con la propia rojez o verdez.[3] De modo que deberíamos esperar que estas relaciones estén cognitivamente representadas y eso es, precisamente, lo que encontramos. Las similitudes y diferencias entre las experiencias corresponden a similitudes y diferencias representadas en la percatación; la geometría de la experiencia corresponde a la geometría de la percatación, etc. Si refinamos la noción de percatación como lo sugerimos más arriba, de forma que los estados de percatación estén siempre acompañados por estados de experiencia, entonces también será plausible una correspondencia estructural en la otra dirección: la estructura representada en la percatación se refleja en la estructura de la experiencia.

Es así que junto con el principio general de que allí donde hay conciencia, hay percatación y viceversa, tenemos un principio más específico: la estructura de la conciencia se refleja en la estructura de la percatación y la estructura de la percatación se refleja en la estructura de la conciencia. Lo denominaré el *principio de coherencia estructural*.[4] Esta es una relación central y sistemática entre la fenomenología y la psicología y, en última instancia, puede traducirse en una relación entre la fenomenología y los procesos físicos subyacentes. Como veremos, es útil de un cierto número de modos.

## 3. Más sobre la noción de percatación

Uno de los proyectos filosóficos más interesantes en el estudio de la conciencia es el de refinar la noción de percatación con el fin de hacerla un correlato psicológico más perfecto de la conciencia. En una definición inicial, la percatación corresponde de un modo imperfecto

con la conciencia, pero la noción puede refinarse para resolver los casos problemáticos. En última instancia, nos gustaría caracterizar un estado psicológico que se correlacione de un modo plausible con la experiencia consciente en general, al menos en una variedad de casos con los que estamos familiarizados.

Inicialmente definí la percatación como el estado en el que alguna información es accesible para la información verbal y el control deliberado de la conducta. Algunas consideraciones acerca de la percatación proposicional en ausencia de experiencia sugirieron modificar esto para requerir un acceso *directo*. Otras modificaciones son posibles. La más obvia es que su disponibilidad para la comunicación verbal no es algo estrictamente necesario para la experiencia consciente, como lo sugieren las consideraciones acerca de la experiencia en los mamíferos, aunque es una buena heurística en los casos en los que el lenguaje está presente. Una sugerencia natural es modificar la definición de la percatación en algo así como *disponibilidad directa para el control global*. Es decir, un sujeto se percata de alguna información cuando ella está directamente disponible para ser utilizada en la dirección de una amplia variedad de procesos conductuales. Esto tiene en cuenta la posibilidad de la experiencia en animales no humanos y también cuadra bien con el criterio de informatividad. En los casos en los que la información es comunicable, esta por lo general está disponible para el control global (por ejemplo, en la dirección deliberada de una amplia variedad de conductas). La implicación inversa no siempre es válida (como lo atestigua el caso de los animales), pero al menos en sujetos que tienen la *capacidad* de informar, la disponibilidad de la información para el control global implica, por lo general, su disponibilidad para la comunicación.

Por supuesto, este proyecto de refinamiento no puede ir más allá, ya que carecemos de un medidor de experiencia con el cual confirmar y refinar de un modo empírico estas hipótesis. No obstante, tenemos una buena idea, basada en el caso de primera persona, sobre los estados en los cuales tenemos experiencias y los estados en los que no, y un análisis de lo que ocurre en esos casos usualmente nos permite caracterizar esos estados en términos funcionales. De modo que una reflexión sobre la relación entre la experiencia y la función en casos familiares nos da una considerable ventaja. También podríamos tratar de refinar empíricamente estas hipótesis por medio de la experimentación de primera persona. Por ejemplo, podemos colocarnos a nosotros mismos en un estado funcional dado y ver qué tipo de experiencia tenemos. Con un poco de ayuda de los principios de homogeneidad y fiabilidad, podemos extraer conclusiones de la investigación de situaciones correlativas en otros.

También existe un papel para la consideración empírica de casos más alejados de uno, por ejemplo, el estudio de qué tipos de experiencias es plausible que tengan sujetos que sufren de ciertas patologías o animales no humanos. Por supuesto, nunca podremos estar completamente seguros acerca de qué experiencias ocurren en esos casos, pero algunas conclusiones son mucho más plausibles que otras. Esos casos actúan como un centro para nuestro razonamiento y una ayuda a la imaginación para formular principios plausibles acerca de la conexión entre la experiencia y la función. Los principios pueden basarse en última instancia en análisis no empíricos, pero la atención a los casos empíricos por lo menos vincula este tipo de razonamiento con el mundo real.

Por ejemplo, la reflexión sobre la atribución de experiencia a los mamíferos concuerda con el criterio refinado que sugerí más arriba. Por lo general, estamos preparados para atribuir experiencia perceptual de un estímulo a los mamíferos en los casos en los que la dirección de la conducta puede hacerse depender de ese estímulo, en especial si esto se exhibe en una variedad de diferentes clases de conductas. Si encontramos que la información acerca de un estímulo sólo puede exhibirse en una reacción conductual única y relativamente menor, podríamos suponer que la información es totalmente inconsciente. A medida que su disponibilidad para ser utilizada se vuelve más difundida, se hace más plausible suponer que es experimentada. De modo que la coherencia entre la conciencia y esta noción de percatación es compatible con los datos de primera persona y con el razonamiento natural que concierne a los casos no humanos.

Existen algunos otros casos problemáticos interesantes para el análisis. Un ejemplo es la *ceguera visual* (descripta por Weiskrantz, 1986). Este es un trastorno que surge de un daño en la corteza visual, en el cual la ruta usual del procesamiento de la información visual está dañada, pero en el que la información visual, sin embargo, parece poder ser procesada de un modo limitado. Los sujetos con ceguera visual no ven nada en ciertas áreas de su campo visual o, por lo menos, eso es lo que dicen. Si colocamos una luz roja o verde en su "área ciega" afirman no ver nada. Pero cuando los *forzamos* a hacer una elección sobre lo que se encuentra en esa área —si es una luz roja o verde, por ejemplo— dan la respuesta correcta con mayor frecuencia que una respuesta errónea. De algún modo "ven" lo que está en el área sin realmente *verlo*.

A veces se propone a la ceguera visual como un caso en el cual la conciencia y el papel funcional vinculado se disocian. Después de todo, en la ceguera visual existe discriminación, categorización e incluso comunicación verbal de algún tipo, pero no parece haber ninguna

experiencia consciente. Si este fuese verdaderamente un caso en el cual el papel funcional y la experiencia estuviesen disociados, ello claramente plantearía problemas al principio de coherencia. Afortunadamente, la conclusión de que esto es un ejemplo de percatación sin conciencia es infundado. Para empezar, no es *obvio* que no haya ninguna experiencia en estos casos; quizás haya una experiencia débil que mantiene una relación inusual con la información verbal. Más al punto, sin embargo, esto está lejos de ser un caso estándar de *percatación*. Hay una gran diferencia entre los papeles funcionales desempeñados aquí y aquellos desempeñados en el caso usual; es precisamente debido a esta diferencia en los papeles funcionales que advertimos que algo anda mal en primer lugar.[5]

En particular, los sujetos con ceguera visual parecen carecer del tipo usual de acceso a la información disponible. Su acceso es curiosamente indirecto, como lo atestigua el hecho de que no está directamente disponible para su comunicación verbal y para el control deliberado de la conducta. La información está disponible para muchos menos procesos de control que la información perceptual estándar; puede volverse disponible para otros procesos, pero sólo mediante métodos inusuales como la incitación y la elección forzada. Por lo tanto, no puede considerarse que esta información esté directamente disponible para el control global, y los sujetos no se percatan verdaderamente de ella en el sentido apropiado. A la ausencia de experiencia le corresponde directamente una ausencia de percatación. También es posible, quizá, que los sujetos con ceguera visual tengan un tipo débil de experiencia, en cuyo caso también podríamos querer decir que tienen un tipo débil de percatación si formulamos las normas de inmediatez y globalidad de la forma apropiada. La descripción de la situación está algo subdeterminada, dado que no tenemos acceso a los hechos en cuestión, pero de cualquier forma es compatible con la coherencia entre la conciencia y la percatación.

En general, este tipo de caso no puede proporcionar evidencia *en contra* de un vínculo entre la organización funcional y la experiencia consciente, ya que nuestras conclusiones acerca de la presencia o ausencia de la conciencia en estos casos se apoyan precisamente sobre bases funcionales. En particular, la evidencia en favor de estados inusuales de conciencia en estos casos patológicos se basa enteramente, por lo general, en la evidencia de estados inusuales de percatación. Estos casos, por lo tanto, no pueden dañar el principio de coherencia; sólo pueden apoyarlo y refinarlo.

Un caso problemático difícil lo proporcionan las experiencias durante el sueño. Es plausible que tengamos experiencias cuando soñamos (aunque véase Dennett, 1978b), pero la informatividad y

cualquier papel en el control de la acción están ausentes, así como la acción también está totalmente ausente. No obstante, estos casos podrían analizarse en términos de su *disponibilidad* para el control global; ocurre sencillamente que los procesos de control pertinentes están en su mayor parte inoperantes. Quizá la información la transforme en el tipo de posición desde la cual puede por lo general usarse con propósitos de control; esta sugerencia está apoyada por la accesibilidad del contenido actual del sueño en un estado de semivigilia. Todavía, entonces, podríamos utilizar el contrafáctico: si la informatividad y el control hubieran estado habilitados (por ejemplo, si la corteza motora hubiese estado funcionando normalmente), entonces la información podría haber desempeñado un papel. Pero esto merece un análisis más cuidadoso, junto con la investigación empírica de lo que realmente ocurre durante el sueño.

Block (1995) presenta algunos casos interesantes en su extensa discusión sobre la distinción entre la conciencia fenoménica y la "conciencia de acceso". En la concepción de Block, un estado es "consciente de acceso" si su contenido puede utilizarse como premisa en el razonamiento, para el control racional de la acción o para el control racional del habla. De modo que la conciencia de acceso corresponde aproximadamente a mi definición inicial de percatación, aunque mi definición no le asigna un papel tan importante a la racionalidad. Block presenta algunos casos en los que las dos variedades de la conciencia podrían estar disociadas. Es instructivo ver cómo un principio de coherencia podría manejarlos.

En la posibilidad de la conciencia de acceso sin conciencia fenoménica, Block apela sólo a casos que son conceptualmente posibles, como los zombis; estos casos no reales claramente no pueden amenazar el principio de coherencia. Menciona la ceguera visual, pero advierte que esta sólo produce conciencia de acceso en un sentido débil. También analiza casos como una "superceguera visual" que es como la ceguera visual excepto que se entrena al sujeto para que tenga un acceso mucho mejor a la información en el área ciega. Existen casos claramente *concebibles* de percatación sin conciencia en las proximidades, pero el propio Block advierte que no hay ninguna razón para creer que esos casos sean reales. Es interesante su observación de que en los ejemplos empíricos más cercanos a un caso de este tipo (un mono descripto en Humphrey, 1992, y un paciente humano descripto en Weiskrantz, 1992; véase también Cowey y Stoerig, 1992) hay razones para creer que existe conciencia fenoménica.

En la conciencia fenoménica sin conciencia de acceso, Block menciona algunos casos reales. Uno es una situación en la cual un sujeto repentinamente se percata del hecho de que hubo un fuerte

ruido de fondo de un taladro funcionando durante algún tiempo. Block sugiere que antes de advertirlo, el sujeto estaba fenomenológicamente consciente pero no tenía conciencia de acceso del ruido del taladro. Utilizando la concepción de percatación que formulé, sin embargo, parece razonable decir que el sujeto se percató todo el tiempo de la perforación. Es plausible que la información pertinente acerca del taladro estuviese *disponible* todo el tiempo; simplemente no *accedió* a ella. De modo que si la conciencia de acceso o la percatación se definen de un modo disposicional, este caso no resulta problemático para un principio de coherencia. Block también menciona un caso en el que una matriz de letras de tres por tres se presenta brevemente a un sujeto (Sperling, 1960). Si se les pide que nombren las letras de la fila superior, los sujetos podrán hacerlo pero no podrán nombrar las letras de las otras filas; lo mismo ocurre con las otras filas. Block argumenta que un sujeto está fenoménicamente consciente de las nueve letras, pero tiene conciencia de acceso de sólo tres letras por vez. No obstante, una vez más, es plausible que la información acerca de las nueve letras estuviese inicialmente disponible; es sólo que únicamente accedió a la información de tres letras, y el mismo proceso de acceso destruyó la accesibilidad del resto de la información. De modo que este caso es también compatible con el principio de coherencia, en una concepción disposicional de la percatación.

Hay muchos otros casos que deberían considerarse. Todo lo que aquí hice fue presentar algunos casos y realizar un breve análisis como ilustración, para dar alguna idea de la forma de un proyecto filosófico interesante. En un análisis más cuidadoso, podríamos intentar imponer restricciones más fuertes sobre exactamente qué tipo de accesibilidad concuerda con la experiencia consciente y precisamente qué tipo de papel de control es apropiado. La concepción de la percatación en términos de la disponibilidad directa para el control global es sólo un inicio. Esta es un área fértil para análisis ulteriores.

## Relación con las teorías funcionalistas de la conciencia*

El proyecto que esbocé puede verse como una búsqueda de un tipo de concepción funcionalista de la conciencia. No es una concepción funcionalista *reductiva*: no dice que el desempeño de algún papel funcional sea todo lo que hay en la conciencia o todo lo que hay que explicar. Más bien, es una concepción *no reductiva*, una que enuncia criterios funcionales de cuándo surge la conciencia. Aun así, existe un sentido en el cual juega en el mismo campo que las concepciones

funcionalistas reductivas: estas también enuncian criterios funcionales de cuándo surge la conciencia junto con sus aseveraciones metafísicas más ambiciosas. Es interesante dejar de lado las diferencias metafísicas y comparar las diversas concepciones en términos de sus criterios funcionales solamente.

Por ejemplo, la propuesta de que la conciencia acompaña a la disponibilidad directa para el control global es reminiscente de la propuesta de Dennett (1993b) de que la conciencia es *celebridad cerebral*: "La conciencia es celebridad cerebral, nada más y nada menos. Son conscientes los contenidos que perseveran, que monopolizan recursos durante un tiempo suficiente como para lograr cierto tipo de efectos típicos y 'sintomáticos', sobre la memoria, sobre el control de la conducta, etcétera" (p. 929).

Dejando de lado el hecho de que Dennett la considera una verdad conceptual, esta se encuentra bastante cerca de nuestra concepción. La diferencia principal es que en mi opinión la conciencia acompaña a la celebridad cerebral *potencial*. No se requiere que un contenido realmente desempeñe un papel de control global para que sea consciente, pero debe estar disponible para poder hacerlo. Esto parece adecuarse mejor a las propiedades de la experiencia. Por ejemplo, experimentamos la periferia de nuestro campo visual, pero la mayor parte del tiempo esta no tiene un gran papel en el control global; pero puede tenerlo si fuera necesario. Muchos de los ruidos que experimentamos pueden pasar sin dejar efectos significativos en la memoria, la conducta, o similares, pero la información *podría* haber tenido efectos. Por supuesto, es posible que Dennett utilice el término "conciencia" en un sentido más fuerte, un sentido según el cual no estamos conscientes de la periferia o de los ruidos (después de todo, Dennett tiene dudas acerca de la propia idea de la experiencia), pero la comparación es interesante de todos modos.

Otra concepción funcionalista es la propuesta por Rosenthal (1996) de que para que un estado sea consciente debe ser el objeto de un pensamiento de orden superior. En el lenguaje que hemos estado utilizando, esto significa que un estado de primer orden es un contenido de la conciencia precisamente cuando existe un juicio de segundo orden acerca de él. Esto es considerablemente más fuerte que mi propuesta, en el mismo modo que la propuesta de Dennett es más fuerte. A primera vista, hay pocas razones para creer que formamos juicios de segundo orden acerca de todas nuestras experiencias, incluyendo las experiencias de cada detalle del campo visual, de los ruidos de fondo, etc. Rosenthal sostiene que los juicios de segundo orden son usualmente inconscientes, por lo cual no advertimos que




*Figura 6.1.* Zippy Cabeza de Piña acerca de las teorías de orden superior de la conciencia. (Reproducido con autorización especial de King Features Syndicate.)

están presentes, pero aun las consideraciones de tercera persona parecen ir en contra de ellos. Todos estos juicios de segundo orden parecen relativamente innecesarios en el diseño de un sistema cognitivo. Podríamos esperar que un sistema tenga la capacidad de formar esos juicios cuando es necesario, como lo hacemos con nuestras experiencias más sobresalientes, pero un sistema con un juicio de segundo orden para cada detalle del campo visual parecería bastante redundante (fig. 6.1).

En la concepción de Rosenthal, los estados conscientes son estados *de los* que somos conscientes. Esto puede tener un tinte de plausibilidad, pero creo que es sólo en el sentido débil (del último capítulo) en el cual estamos *familiarizados* con todas nuestras experiencias.[6] No es claro en absoluto que la mayoría de nuestras experiencias sean objetos de nuestros pensamientos. Suponer que existen dos estados cognitivos separados para cada detalle de la experiencia, un juicio de primer orden y uno de segundo orden, lleva a una imagen atiborrada de la mente. Es difícil ver por qué la evolución se molestaría en incorporar los juicios de segundo orden en general, cuando una simple disponibilidad para el control global serviría igualmente bien a sus propósitos. Es mejor interpretar la teoría de Rosenthal como una concepción de la conciencia introspectiva, aunque él la formula como una concepción de la conciencia en el sentido de "cómo es".

Es útil dividir las concepciones funcionales de la conciencia en variedades *de primer orden* y *de segundo orden*. En las variedades de segundo orden (que incluyen la concepción de pensamiento de orden superior de Rosenthal, además de la concepción de percepción de orden superior de Lycan [1995] y otras), lo fundamental para la

conciencia es la presencia de algún estado cognitivo de segundo orden. En las teorías de primer orden, sólo se requiere un estado cognitivo de primer orden, con algunas restricciones sobre el papel que desempeña. Las teorías de segundo orden pueden ofrecer una buena concepción de la introspección o de la conciencia reflexiva, pero las teorías de primer orden parecen estar mucho más estrechamente vinculadas a la experiencia consciente.[7]

Por supuesto, no *todos* los estados cognitivos de primer orden corresponden a experiencias conscientes; puede haber juicios de primer orden acerca del mundo que no correspondan a ninguna experiencia en absoluto. Por lo tanto, necesitamos un componente adicional en una teoría de este tipo para distinguir las clases apropiadas de estados de primer orden. El modo obvio de hacerlo es restringir el *papel* de esos estados. Esto es lo que yo hice cuando sugerí que los juicios de primer orden pertinentes son precisamente aquellos que están directamente disponibles para el control global. Como veremos, otras concepciones de primer orden sugieren restricciones relacionadas. Podríamos argumentar que la causalidad de un pensamiento de orden superior es sólo otra restricción de este tipo; el problema es que la restricción parece ser excesivamente fuerte.

Una interesante propuesta intermedia es que un estado consciente corresponde a un juicio de primer orden que tiene la *capacidad* de causar un juicio de segundo orden acerca de él.[8] Esto evita el atiborramiento de la propuesta previa y tiene un elemento de plausibilidad. No es tan diferente de la noción de que un estado consciente corresponde a un juicio de primer orden que está disponible para el control global: podemos suponer que la disponibilidad para el control global y la disponibilidad para el juicio de segundo orden van de la mano gran parte del tiempo. Sin embargo, una propuesta de este tipo puede tambalear cuando se trata de sistemas que, como los bebés y los animales, supuestamente tienen experiencias pero carecen de la capacidad para realizar juicios de segundo orden; esto parece requerir una mayor sofisticación conceptual de la que puede necesitarse para la posesión de experiencias. Si esto es así, la caracterización en términos de disponibilidad para el control global es superior.

Es probable que cualquier concepción funcionalista de primer orden invoque una restricción que involucra algún tipo de disponibilidad. Un examen de las concepciones reductivas existentes lo corrobora. Por ejemplo, Kirk (1992) sugiere que la conciencia perceptual requiere que la información ingresante esté "presente" para los principales procesos de toma de decisiones de un sistema, y Kirk (1994) sugiere que se requiere información "directamente activa". La sugerencia de Dretske (1995) de que la experiencia es información

que está representada *para* un sistema también tiene este sabor, como también la sugerencia de Tye (1995) de que la información debe estar "preparada" de un modo apropiado para el procesamiento cognitivo. Es probable que estas sugerencias puedan reconciliarse entre sí sin demasiadas dificultades. Todas parecen tener el propósito de expresar una idea similar.

En cualquier caso, es interesante que un no reduccionista acerca de la conciencia no necesite considerar las cuestiones entre las diversas concepciones funcionalistas de primer orden y de segundo orden como una guerra mutuamente destructiva entre teorías condenadas. Aunque estas concepciones no pueden explicar la conciencia, son, sin embargo, relativamente significativas como teorías candidatas de la base cognitiva de la conciencia, y algunas de ellas tienen aquí más éxito que otras. Incluso un dualista de propiedades puede reconocer un elemento de verdad en ellas y puede asignarle alguna importancia a sus diferencias.

## Juicios de primer orden y registros de primer orden*

Argumenté que los estados de experiencia corresponden directamente a estados cognitivos subyacentes que denominé juicios de primer orden. Pero como hice notar antes, y como Dretske (1995) puso de relieve, podría ser engañoso llamar a estos estados *juicios*. Los juicios, recordemos, fueron definidos originalmente como algo bastante similar a las creencias (con la estipulación de que se excluye cualquier elemento fenoménico). No obstante, aunque es razonable suponer que existe un estado representacional que corresponde a todos los detalles en un campo visual experimentado, no es claro que el sujeto tenga creencias acerca de todos esos detalles. El contenido de la periferia de mi campo visual, por ejemplo, podría ser algo acerca de lo cual yo no tengo creencias en un sentido u otro, al menos hasta que le presto atención. Sin embargo, aun en ausencia de creencias, existe algún tipo de estado cognitivo que contiene la información pertinente, ya que ella está, al menos, *disponible*.

Podríamos simplemente estipular que utilizamos el término "juicio" de un modo más amplio para cubrir este tipo de estados cognitivos además de las creencias explícitas. Después de todo, es plausible que las representaciones de la periferia del campo visual puedan considerarse "microjuicios" o juicios implícitos hechos por procesos dentro de los sistemas cognitivos, aunque no sean juicios de toda la persona. Pero, probablemente sea mejor evitar confusiones en esta cuestión e introducir un término más amplio para los estados

representacionales que no son necesariamente juicios. Utilizaré el término "registros" para este propósito. El contenido cognitivo de los estados perceptuales, por ejemplo, estará contenido en los registros de primer orden en lugar de en los juicios de primer orden. Un registro de primer orden no tiene por qué ser un estado reconocido por el sujeto, pero es, no obstante, un estado con contenido que está disponible al sujeto y que desempeña un papel en el sistema cognitivo.

Los registros de primer orden pueden incluso, de vez en cuando, ser contradichos por los juicios de primer orden. Las ilusiones ópticas son un claro ejemplo: un sujeto puede saber que dos objetos tienen el mismo tamaño, pero la percepción puede representarlos, de todas maneras, con tamaños diferentes. Dretske ofrece otro ejemplo. Usted me muestra siete dedos y yo veo los siete. Pero no tengo tiempo para contarlos y erróneamente supongo que veo ocho dedos. De modo que *juzgo* que hay ocho dedos ante mí, pero mi experiencia fenoménica es de siete dedos. El juicio, por lo tanto, no es directamente paralelo a la fenomenología. No obstante, en algún lugar dentro del sistema perceptual, la información visual de siete dedos está representada y se encuentra disponible para los sistemas subsiguientes. Es a esta representación previa a lo que denomino un registro de primer orden. Podemos pensar en los registros de primer orden como el producto inmediato de procesos perceptuales e introspectivos, antes de que sean racionalmente integrados en un todo coherente.[9]

El contenido de la percatación, entonces, estará constituido, estrictamente hablando, por registros de primer orden en lugar de juicios de primer orden. En particular, el contenido de la percatación consistirá aproximadamente en los registros de primer orden que están directamente disponibles para su utilización en el control global. Así definido, el contenido de la percatación corresponde directamente al contenido de la conciencia. Por supuesto, habrá algunos registros de primer orden que caigan fuera del contenido de la percatación, como en los estados de percepción subliminal, por ejemplo. Como con los juicios, podemos hablar de registros *fenoménicos* de primer orden para distinguir los registros que corresponden a experiencias de aquellos que no. Siempre me ocuparé de la primera clase, sin embargo, de modo que por lo general hablaré simplemente de "registros de primer orden" y dejaré implícito el modificador.

(Probablemente sea mejor considerar el contenido representacional de un registro de primer orden como contenido *no conceptual*, como lo es el contenido paralelo de la experiencia. Analizaré esta y

algunas otras cuestiones acerca del contenido de la percatación y la experiencia en una nota aparte.)[10]

## 4. El papel explicativo de los principios de coherencia

Los principios de coherencia que esbocé no son sólo ornamentos metafísicos. Pueden tener un papel fundamental en el trabajo empírico sobre la experiencia consciente. Cualquier estudio empírico de la conciencia requiere algún razonamiento preexperimental para al menos despegar, con el fin de extraer conclusiones acerca de la experiencia consciente sobre la base de los datos físicos. Los principios de coherencia proporcionan la base necesaria. Con ellos en su lugar, existe una base metodológica para la investigación empírica de la experiencia consciente en un cierto número de áreas. Ya se realizan muchos trabajos en este dominio; los principios de coherencia simplemente ponen en descubierto los supuestos que subyacen a esos trabajos.

Existen por lo menos tres grandes proyectos en los que estos principios podrían tener un papel explicativo. Primero, el principio de coherencia estructural nos puede ayudar en el proyecto de utilizar hechos acerca del procesamiento físico para contribuir a la explicación de la estructura de tipos específicos de experiencias. Segundo, la coherencia entre la conciencia y la percatación actúa como una especie de palanca epistémica que les permite a los investigadores inferir conclusiones acerca de la experiencia a partir de datos de tercera persona. Y tercero, la coherencia entre la conciencia y la percatación puede servir como principio de fondo en la búsqueda de los correlatos físicos de la conciencia. Analizaré cada uno de ellos.

El primero de estos proporciona el ejemplo más claro en la práctica contemporánea. Es común ver cómo se utilizan trabajos empíricos sobre los procesos neurobiológicos y cognitivos para arrojar luz sobre las características estructurales de la experiencia. Como ya lo analicé, por ejemplo, un estudio de los procesos subyacentes en la visión del color es útil para ayudar a explicar la estructura del espacio fenoménico del color. De un modo similar, el estudio de mapas topográficos de la corteza visual ayuda a arrojar luz sobre la estructura del campo visual fenoménico, y el estudio del procesamiento en la corteza auditiva nos ayuda a comprender muchos aspectos estructurales de las experiencias auditivas (relaciones de tono y aspectos direccionales, por ejemplo). Algo similar se aplica a muchos otros dominios fenoménicos.

Podríamos preguntarnos cómo cualquier historia acerca de los procesos físicos podría utilizarse para arrojar luz sobre las características de la experiencia, dado lo que ya dije sobre la imposibilidad de la explicación reductiva. El principio de la coherencia estructural nos permite comprender lo que ocurre. En esencia, este principio se utiliza como un *supuesto general* que proporciona un puente entre las características de los procesos físicos y las características de la experiencia. Si damos por sentado la coherencia entre la estructura de la conciencia y la estructura de la percatación, entonces para explicar algún aspecto específico de la primera, sólo necesitamos explicar el aspecto correspondiente de la segunda. El principio puente hace el resto del trabajo.

En el caso del color, por ejemplo, lo que ocurre es que una historia acerca de los procesos físicos nos da una concepción reductiva de la estructura de la percatación al explicar las similitudes y diferencias significativas entre los estímulos visuales que el sistema de color procesa y hace disponibles a los sistemas subsiguientes. Una vez que tenemos esta concepción de la estructura de la percatación del color, el principio de coherencia nos dice que esta estructura se reflejará en la estructura de la experiencia del color. De modo que si se acepta el principio de coherencia, una concepción funcional del procesamiento visual sirve como una concepción indirecta de la estructura del espacio de color fenoménico. El mismo método puede utilizarse para explicar muchas otras características de la experiencia.

Algunas personas se han sentido lo suficientemente impresionadas por la coherencia entre la estructura en la conciencia y en la cognición como para sugerir que esto es todo lo que necesitamos para una explicación física de la conciencia. Van Gulick (1993), por ejemplo, hace notar el hecho de que la estructura de nuestro espacio de color corresponde directamente a una estructura que está representada en el procesamiento visual, y sugiere que esto cierra la "brecha explicativa" al proporcionar una explicación funcional de la sensación del color. Clark (1993) dedica todo un libro a esta estrategia, argumentando que las cualidades sensoriales pueden explicarse completamente dando cuenta de las relaciones de similitud y diferencias dentro de los espacios de cualidades.

Si lo que dije antes es correcto, estas afirmaciones son un poco fuertes. Primero, este método no explica la naturaleza *intrínseca* de una experiencia de color, como lo muestra la posibilidad de una inversión del espectro que preserve la estructura. Cuanto más, explica la estructura relacional *entre* esas experiencias o entre las partes de una experiencia compleja; de modo que se requiere algo más para una

concepción completa de la conciencia. Segundo y más importante, ninguna concepción de la estructura de la percatación explica en absoluto por qué existe una experiencia acompañante, precisamente porque no puede explicar por qué el principio de coherencia estructural es válido en primer lugar. Al tomar el principio como un supuesto general ya nos hemos movido más allá de la explicación *reductiva*: el principio simplemente supone la existencia de la conciencia, y no hace nada por explicarla. Esto puede considerarse una especie de explicación *no reductiva*, que da por sentada la existencia de la conciencia e intenta explicar algunas de sus propiedades.

Dentro de estos límites, el principio de coherencia estructural proporciona una relación explicativa enormemente útil entre lo físico y lo fenoménico. Si queremos explicar alguna estructura aparente en un dominio fenoménico —digamos, las relaciones que encontramos entre nuestras experiencias de los acordes musicales—, entonces podemos investigar la organización funcional del dominio psicológico correspondiente tomando ventaja de los descubrimientos de la ciencia cognitiva y la neurociencia para explicar reductivamente la estructura de la percatación en ese dominio. Al hacerlo de esta manera explicamos la estructura del domino fenoménico, módulo la contribución del principio de coherencia estructural. Debido a nuestra apelación a este principio no habremos explicado la propia conciencia, pero de todas formas habremos explicado gran parte de lo que es especial acerca de un dominio fenoménico *particular*.

De esta forma, el principio de coherencia estructural puede servir de columna vertebral de un proyecto que Crick y Koch[11] denominan "la historia natural de los qualia". Aunque la neurociencia no puede explicar la existencia de la experiencia, sí puede explicar un vasto número de hechos acerca de ella. La neurociencia puede indirectamente explicar las relaciones de similitud y diferencias entre experiencias; la geometría de los espacios experienciales, como el espacio del gusto y el espacio del color; la estructura detallada de campos experienciales, como el campo visual; la localización percibida asociada con experiencias dentro de ese campo; la intensidad de las experiencias; la duración de las mismas; asociaciones entre las experiencias; y mucho más. Como Crick y Koch lo expresan, la neurociencia puede formular una concepción de todas las características de la experiencia que son objetivamente comunicables. La propia comunicabilidad de esas características implica que se reflejan en características físicas del sistema y ciertamente en características de la percatación. La coherencia estructural entre la conciencia y la percatación es la base implícita o explícita sobre la cual se apoya

este tipo de explicación. (Una base de este tipo es particularmente crucial en el campo de la psicofísica, como expongo en una nota agregada.)[12]

Utilizando estos métodos, ¡podríamos incluso obtener alguna comprensión de cómo es ser un murciélago! La organización funcional nos puede decir mucho acerca del tipo de información a la que un murciélago tiene acceso— los tipos de discriminaciones que puede hacer, los modos como categoriza las cosas, las propiedades más sobresalientes en su campo perceptual, etc. —y acerca del modo como la utiliza. Eventualmente deberíamos ser capaces de construir una imagen detallada acerca de la estructura de la percatación en el sistema cognitivo de un murciélago. Mediante el principio de coherencia estructural, tendremos entonces una buena idea acerca de la estructura de las experiencias del murciélago. No sabremos *todo* acerca de cómo es ser un murciélago —no tendremos una clara concepción de la naturaleza intrínseca de las experiencias, por ejemplo— pero sabremos bastante. Un artículo interesante de Akins (1993) acerca de la vida mental de los murciélagos puede leerse como una contribución a este proyecto.

De un modo similar, el libro de Cheney y Seyfarth (1990) *How Monkeys See the World* es formulado como respuesta a una pregunta como la que hicimos acerca de los murciélagos; esto nos lleva dentro de la mente de otras especies. De hecho, el trabajo utiliza el principio de coherencia estructural como un supuesto de base en todo momento; ofrece una concepción de ciertos procesos funcionales y de la estructura de la percatación que implican, y nos invita a inferir una estructura correspondiente de la experiencia. Por supuesto esto no responde a la real preocupación de Nagel, por las razones usuales, pero es, de todas formas, un logro notable. No necesitamos enfrentar el misterio último de la conciencia cada vez que queremos explicar un domino fenoménico específico.

## Los principios de coherencia como palancas epistémicas

Los investigadores empíricos en la neurociencia, la psicología, la etología, y campos relacionados a veces quieren hacer aseveraciones acerca de la presencia de la experiencia consciente en un sistema. Aunque la conciencia muy frecuentemente es dejada de lado en estos campos, existe un cuerpo de trabajos de cierta importancia en los que se extraen conclusiones acerca de la experiencia consciente a partir de los resultados empíricos. ¿Cómo es esto posible, dadas las dificultades en la observación directa de la experiencia? Si todo

lo que puede observarse son procesos físicos, ¿qué justifica cualquier conclusión?

La respuesta debe ser que cada vez que se extraen conclusiones acerca de la experiencia a partir de resultados empíricos, un principio puente que vincula los procesos físicos con la experiencia es el que realiza el trabajo. Un principio puente definirá un *criterio* para la presencia de la conciencia en un sistema, un criterio que se aplica en el nivel físico. Un principio de este tipo actuará como *una palanca epistémica* que permite pasar del conocimiento de los procesos físicos al conocimiento acerca de la experiencia. La palanca epistémica no es ella misma experimentalmente verificable, al menos desde el punto de vista de tercera persona; en cambio, actúa como un tipo de supuesto general previo. Estos supuestos no siempre se hacen explícitos, pero son el único modo como este tipo de trabajo logra alguna base en la experiencia consciente.

Los principios puente son tan cruciales aquí que es importante considerarlos explícitamente. Existe un sentido en el cual cualquiera que recurra a un principio puente —es decir, cualquiera que extraiga conclusiones acerca de la experiencia a partir de las observaciones externas— hace "filosofía", ya que los principios puente no son ellos mismos conclusiones experimentales. Estos principios deben basarse en consideraciones a partir del caso de primera persona y en principios generales de plausibilidad. Preceden efectivamente a cualquier resultado experimental, ya que son los propios principios los que nos dicen cómo interpretar los resultados. Por supuesto, existen supuestos *a priori* involucrados en cualquier empresa experimental, pero aquí estos tienen un papel inusualmente significativo. Por lo tanto, es importante justificar esos supuestos tan bien como podamos, mediante un análisis cuidadoso. Esta es una manera de interpretar el proyecto en el que me he embarcado en este capítulo.

El principio puente que yo recomendé es el de la coherencia entre la conciencia y la percatación: cuando un sistema se percata de alguna información, en el sentido de que la información está directamente disponible para el control global, entonces la información es consciente. Sospecho que si realizáramos un estudio cuidadoso de los principios puente utilizados por los investigadores empíricos y por aquellos que interpretan la investigación empírica, casi todos esos principios serían compatibles con este y serían derivables de él. El principio puente más común, por supuesto, es el uso de la informatividad como criterio para la experiencia: al menos en un sistema que utilice un lenguaje, se considera que, por lo general, la información es consciente si es comunicable. La informatividad es una versión de la percatación

—cuando la información es comunicable, está siempre disponible para el control— de modo que este criterio claramente concuerda con el principio de coherencia, aunque es más limitado en su alcance.

Ocasionalmente se utilizan también otros criterios; a veces los investigadores quieren hacer aseveraciones acerca de la experiencia en animales sin lenguaje o en seres humanos cuyos mecanismos de comunicación no funcionan normalmente. En estos casos, usualmente se considera que el mejor signo de la experiencia es un efecto fuerte de alguna información en el control de la conducta. Por ejemplo, Logothetis y Schall (1989) presentan su trabajo como el aislamiento de los "correlatos neuronales de la percepción visual subjetiva" en monos. Aquí, se interpreta que un mono tiene una experiencia perceptual de un objeto en movimiento en su ambiente cuando puede fiablemente hacer un movimiento ocular o presionar una palanca en respuesta a ese movimiento. Una vez más, esto concuerda perfectamente con el criterio provisto por la percatación o con la disponibilidad directa para el control global.

Algunos podrían encontrar que una apelación a principios puente preexperimentales es perturbadora para una ciencia experimental; ciertamente, la necesidad de estas palancas epistémicas podría ser la razón de que esas ciencias se hayan mantenido con tanta frecuencia apartadas de la conciencia. Sin embargo, este es el bote en el que nos encontramos, y las conclusiones extraídas sobre la base de estos principios son mejores que ninguna conclusión. Tiene sentido explicitar los principios relevantes, sin embargo, y que se los justifique mediante un análisis cuidadoso, en lugar de ocultarlos debajo de la alfombra, De este modo podrá clarificarse el razonamiento subyacente que lleva a conclusiones empíricas acerca de la experiencia consciente.

## Los correlatos físicos de la conciencia

¿Cuáles son los *correlatos* neuronales y de procesamiento de la información de la conciencia? Esta es una de las preguntas fundamentales acerca de la conciencia, de la que frecuentemente se supone que la investigación empírica se ocupa. Se formularon diversas hipótesis empíricas. Por ejemplo, Crick y Koch (1990) formularon la hipótesis de que ciertas oscilaciones de 40 hertz en la corteza son el correlato neuronal de la experiencia. Se puede interpretar que Baars (1988) sugiere que un espacio de trabajo global es la base del procesamiento de la información de la experiencia, donde el contenido de esta corresponde directamente al contenido del espacio de trabajo.

Farah (1994) argumenta que la conciencia está asociada a las representaciones de "alta calidad" en el cerebro. Libet (1993) formula una teoría de "tiempo de persistencia" neuronal, en la que se asocia la conciencia asociada a las actividades neuronales que persisten durante un tiempo suficiente, con una duración mínima de alrededor de 500 milisegundos. Han existido otras numerosas propuestas en una vena similar.

La coherencia entre la conciencia y la percatación proporciona un modo natural de comprender gran parte de estos trabajos. Es sorprendente que cada uno de estos candidatos sea él mismo un candidato plausible para desempeñar un papel en la facilitación de la *percatación*: la disponibilidad directa para el control global. Las oscilaciones de Crick y Koch se formulan debido al papel que podrían desempeñar en vincular la información y colocarla en la memoria de trabajo; y, por supuesto, la memoria de trabajo es sólo un sistema mediante el cual el contenido puede hacerse disponible para el control. La noción de actividad neuronal temporalmente extendida de Libet puede ser pertinente porque ese tipo de actividad tiene los sólidos efectos generalizados sobre el sistema cognitivo requeridos para la percatación. Lo mismo ocurre con las representaciones de "alta calidad" de Farah; es posible que las representaciones de "baja calidad" puedan no ser capaces de impregnar el funcionamiento cognitivo del modo apropiado. En el caso del espacio de trabajo global de Baars, el vínculo es el más claro de todos: el autor formula el espacio de trabajo precisamente en virtud de su papel en la mediación del acceso y el control globales.

Una interpretación deflacionaria de lo que ocurre aquí sería que estos investigadores simplemente *quieren decir* percatación cuando dicen "conciencia", de modo que este punto en común no es sorprendente. No obstante, creo que es evidente a partir del contexto que la mayoría de ellos —al menos Crick y Koch, Farah y probablemente Libet— hablan de la conciencia en su sentido fenoménico completo e intentan aislar sus correlatos físicos. Todos ellos hacen observaciones que sugieren que aceptarían una distinción *conceptual* entre la conciencia y la percatación tal como la defino aquí.

Una interpretación más interesante es suponer que estos investigadores hablan sobre la conciencia en el sentido fenoménico, y observar que todas sus propuestas son compatibles con el principio puente general de la coherencia entre la conciencia y la percatación. Estas hipótesis pueden ser *derivables* a partir del principio de coherencia, junto con los resultados empíricos pertinentes. Digamos que aceptamos el principio de coherencia como un supuesto general,

lo que significa que aceptamos que la experiencia está directamente asociada a la disponibilidad directa para el control global. Si los resultados empíricos sugieren que en una especie particular (como el *Homo sapiens*), las oscilaciones de 40 hertz tienen una función en la disponibilidad global, entonces tenemos razones para creer que las oscilaciones son un correlato de la experiencia en esa especie. Si los resultados sugieren que la actividad temporalmente extendida sirve a la disponibilidad global, entonces tenemos razones para creer que ese tipo de actividad es un correlato de la experiencia. Y así sucesivamente.

Por supuesto, más de una de estas hipótesis podría ser correcta. Quizá tanto las oscilaciones como la actividad temporalmente extendida tengan una función en la disponibilidad global en diferentes instancias, o quizá desempeñan simultáneamente un papel en diferentes etapas del proceso de acceso/control. Quizá las oscilaciones sirvan a las representaciones de alta calidad en el espacio de trabajo global. Esas son cuestiones empíricas. Pero las hipótesis también podrían resultar falsas. Quizá resulte que las oscilaciones no tienen ningún papel especial en el control global y, en cambio, estén sólo involucradas en las operaciones periféricas. Quizá solamente tienen efectos muy limitados sobre los procesos posteriores y sobre la conducta.

Lo que es notable es que *si* tuviésemos razones para creer que las oscilaciones están disociadas de la percatación de ese modo, también tendríamos razones para creer que están disociadas de la experiencia. Si resultase que las oscilaciones no tienen ninguna relación especial con la informatividad y la conciencia, por ejemplo, el fundamento de la hipótesis de correlación quedaría eliminado. Después de todo, no tenemos evidencia independiente en favor de la hipótesis: toda nuestra evidencia proviene del vínculo con la informatividad y la percatación. Debido a que no poseemos un "medidor de la experiencia", siempre debemos basarnos en criterios indirectos, y los criterios de informatividad y percatación parecen ser lo mejor que tenemos. Se deduce que sólo podemos tener evidencia empírica de un vínculo entre un proceso $N$ y la conciencia si ya tenemos evidencia en favor de un vínculo entre $N$ y la percatación.

Esto sugiere una metodología clara para encontrar correlatos físicos de la experiencia. Las consideraciones preexperimentales sugieren que el correlato del procesamiento básico de la conciencia es la percatación, o disponibilidad global. Los resultados empíricos sugieren que los estados físicos que desempeñan un papel en la percatación en una especie determinada son de un

cierto tipo *N*. El principio puente previo y los resultados empíricos se combinan para sugerir que en esta especie, *N* es un correlato físico de la conciencia.

(Es interesante que Dennett (1993b) sugiere casi la misma metodología. Sugiere que las propuestas específicas acerca de la base de la conciencia, como la propuesta de la oscilación, son plausibles precisamente en la medida en que los procesos pertinentes tienen un papel en garantizar la "celebridad cerebral". Puedo estar de acuerdo en casi todo esto, excepto que yo sólo pido celebridad cerebral *potencial* y creo que el vínculo entre la conciencia y la celebridad es un principio nomológico en lugar de una verdad conceptual. *Cualquiera* que investigue empíricamente la conciencia necesitará un principio puente no empírico para interpretar los resultados físicos en términos de la experiencia consciente. Para el reduccionista este será una verdad conceptual, y para el dualista de propiedades será un principio nomológico basado en consideraciones de primera persona y un análisis de plausibilidad. Pero muchos de los puntos que formulo aquí se aplican de todas maneras.)

Parece natural decir que la correlación *central* entre el procesamiento físico y la experiencia es la coherencia entre la conciencia y la percatación. Lo que da lugar *directamente* a la experiencia no son oscilaciones o actividades temporalmente extendidas o representaciones de alta calidad, sino el proceso de disponibilidad directa para el control global. Cualquier estado físico más específico calificará como correlato sólo en tanto desempeñe un papel en la disponibilidad global; de modo que las correlaciones más específicas se derivan de la correlación general.

Puede haber muchos correlatos de este tipo. Por ejemplo, diferentes tipos de procesos físicos pueden servir a la disponibilidad en diferentes modalidades. También podría haber diferentes correlatos en diferentes etapas del camino de procesamiento; aun si tan sólo tomamos la experiencia visual, puede haber un tipo de correlato en la corteza visual y otro en áreas ulteriores del circuito. Por supuesto, no existe ninguna garantía de que incluso dentro de una modalidad particular, haya un correlato neuronal de algún tipo simple. Quizá no exista ninguna caracterización simple de los procesos en la corteza visual que tienen una función en la experiencia; podría ocurrir que estos se agrupen *sólo* por su papel funcional (esto es, por el hecho de que sirvan a la percatación). Pero, al menos podemos esperar que pueda existir una caracterización más directa: si no en las áreas corticales sensoriales, entonces quizás en algún punto posterior del camino de procesamiento.

También podríamos encontrar correlatos en niveles más elevados que el neuronal. En la construcción de un modelo cognitivo o computacional apropiado, podríamos ser capaces de descubrir algún modo de caracterizar en términos de procesamiento de información a aquellas entidades que son responsables de la percatación y que, por lo tanto, subyacen a la experiencia. Por supuesto la caracterización más simple de este tipo es tautológica: los procesos que sirven a la percatación son aquellos responsables del acceso global, el control y la comunicación verbal. Pero un modelo cognitivo sustancial podría ofrecernos caracterizaciones menos tautológicas. Quizá podríamos ser capaces de caracterizar en términos relativamente locales el tipo de información que posee un papel global significativo, por ejemplo, dado el diseño global del sistema. Un ejemplo es la sugerencia de Shallice (1972) de que el contenido de la conciencia corresponde al contenido de las "entradas selectoras" a ciertos sistemas de acción. De acuerdo con el diseño del modelo, las entradas selectoras determinan qué sistemas de acción se vuelven "dominantes" o desempeñan un papel en el control global. Si esto es así, las entradas selectoras facilitan la percatación y son un correlato plausible de la experiencia consciente.

Otros correlatos del procesamiento de información caen en algún lugar entre lo tautológico y lo no tautológico. Un ejemplo es el espacio de trabajo global de Baars. Podríamos *definir* el espacio de trabajo global como una especie de área "virtual", que corresponde precisamente a aquellos contenidos que están ampliamente diseminados; de ser así el contenido del espacio de trabajo es, casi por definición, el contenido de la percatación. El modelo de Baars tiene más consecuencias empíricas que esto: él propone que el espacio de trabajo es un sólo sistema unificado (uno que podemos localizar al menos en términos de procesamiento de la información) en el cual la información está integrada y diseminada. Esto podría resultar falso, de modo que la propuesta tiene sustancia empírica que el trabajo experimental podría confirmar o rechazar. Pero la caracterización del espacio de trabajo todavía está lo suficientemente cerca de la caracterización de la percatación como para explicar el leve matiz *a priori* que la propuesta conserva; a veces, parece que casi cualquier resultado empírico puede hacerse compatible con un marco de este tipo. (Por supuesto Baars hace aseveraciones más específicas acerca de las operaciones del espacio de trabajo, y estas aseveraciones tienen una sustancia empírica significativa.) La propuesta de "representación de alta calidad" de Farah tiene también un leve hálito de tautología, aunque esto depende de cómo se definan las representaciones de alta calidad: si ser una representación de "alta calidad" es sólo poder

desempeñar un papel global significativo, entonces el matiz *a priori* es fuerte, pero si se define en términos del modo como se forma una representación de esta clase, el matiz es mucho más débil.

Aun en el nivel cognitivo, no hay ninguna razón especial para creer que haya un solo mecanismo aislable subyacente a la experiencia. Schacter (1989) sugiere que podría haber un solo mecanismo, algo así como un módulo, pero esta es sólo una manera como las cosas podrían resultar. *Podría* ocurrir que un papel en el control global esté siempre facilitado por algún mecanismo central (como el espacio de trabajo global de Baars) pero, a primera vista, es igualmente probable que procesos de muchos tipos diferentes sean responsables en distintos momentos de asegurar la disponibilidad apropiada, incluso dentro de una sola especie o un solo sujeto.

A veces las personas quieren obtener conclusiones más fuertes acerca de los correlatos físicos que los que sugerí. Por ejemplo, si encontramos que las oscilaciones de 40 hertz son la base de la experiencia en casos familiares, ¿por qué no hipotetizar que las oscilaciones de 40 hertz son la base *última* de la experiencia? ¿Podría ocurrir que esas oscilaciones den origen a la experiencia aun cuando no estén asociadas a la percatación, y que sistemas con un funcionamiento apropiado pero sin las oscilaciones carezcan de experiencia? Una conclusión de este tipo no tendría justificación, sin embargo. Las oscilaciones de 40 hertz fueron consideradas significativas *debido* a su asociación con la percatación; no tenemos ninguna razón para creer que cuando no tienen ese papel, haya algo especial en ellas. ¡No hay ninguna razón para creer que oscilaciones de 40 hertz en una probeta deberían dar origen a experiencias como las mías! Aun en casos intermedios, como los de animales o sistemas anestesiados, sería peligroso inferir cualquier cosa acerca de la experiencia a partir de la presencia de las oscilaciones, excepto en la medida en que su presencia nos dé razones para creer que existe algún tipo de percatación.

En general, no podemos esperar que estos métodos empíricos produzcan principios psicofísicos universales incompatibles con aquellos que utilizamos como punto de partida. Pueden producir principios más *específicos*, cuando se aplican a especies determinadas, pero estos se derivarán mediante una aplicación directa de los principios puente preexistentes. Debido a que los principios preexistentes soportan toda la carga de la extracción de conclusiones acerca de la experiencia a partir de los datos físicos, es imposible que estos apoyen una conclusión que contradiga los principios.

No deberíamos entonces esperar que la búsqueda de un correlato neuronal de la conciencia nos conduzca al Santo Grial de una teoría

universal. Podemos esperar que sea útil para ayudarnos a comprender la conciencia en casos específicos, como el humano: aprender más acerca de los procesos que subyacen en la percatación nos ayudará a comprender la estructura y la dinámica de la conciencia, por ejemplo. Pero, dado que sostienen el puente entre los procesos físicos y la experiencia consciente, los principios preexperimentales de coherencia siempre tendrán un papel fundamental.

## 5. La coherencia como una ley psicofísica

Hasta ahora hemos considerado a la coherencia principalmente dentro de un conjunto de casos relativamente familiares que involucran a los seres humanos y a otros sistemas biológicos. Pero es natural suponer que estos principios de coherencia pueden tener la condición de leyes universales. Si la conciencia está siempre acompañada por la percatación y viceversa, en mi propio caso y en el caso de todos los seres humanos, nos vemos llevados a sospechar que algo sistemático sucede. Ciertamente existe una correlación legaliforme en los casos familiares. Por lo tanto, podemos formular la hipótesis de que esta coherencia es una ley de la naturaleza: en cualquier sistema, la conciencia estará acompañada por la percatación y viceversa. Lo mismo ocurre con el principio de la coherencia estructural. La notable correlación entre la estructura de la conciencia y la estructura de la percatación parece demasiado específica como para que sea accidental. Es natural inferir una ley subyacente: para cualquier sistema, en cualquier lugar en el espacio y el tiempo, la estructura de la conciencia reflejará y será reflejada por la estructura de la percatación.

Leyes como estas harían una contribución significativa a una teoría de la conciencia. Hasta ahora, todo lo que sabemos es que la conciencia surge de alguna manera de lo físico, pero no sabemos en virtud de qué propiedades físicas lo hace; esto es, no sabemos qué propiedades son parte del lado físico de la conexión. Con las leyes de coherencia, tenemos una respuesta parcial: la conciencia surge en virtud de la organización funcional asociada a la conciencia. Podemos incluso llegar a una comprensión relativamente específica de partes de la relación de superveniencia en virtud del principio de coherencia estructural: no sólo la conciencia surge de la percatación, también su estructura está determinada por la estructura de aquella.

Por supuesto, es probable que esta ley no sea una ley psicofísica *fundamental*. Las leyes fundamentales conectan propiedades más básicas o al menos definidas de un modo más nítido que una construcción de alto nivel como la "percatación". Pero no todas las leyes son leyes fundamentales. Puede ocurrir, incluso, que los principios de

coherencia no sean leyes *estrictas*; puede haber excepciones en las proximidades, especialmente dada la naturaleza subdeterminada del concepto de percatación. No obstante, aunque estas leyes no sean ni fundamentales ni estrictas, proporcionan una fuerte restricción que cualquier ley psicofísica fundamental debe satisfacer. Una teoría de la conciencia propuesta que no tenga como consecuencia los principios de coherencia estará en problemas. Recíprocamente, si una ley psicofísica fundamental propuesta es simple, está bien motivada y tiene los principios de coherencia como consecuencia, entonces estas son buenas razones para aceptarla.

¿Cuáles son, entonces, las razones para aceptar los principios de coherencia como leyes? La evidencia básica proviene de las correlaciones en casos familiares: en última instancia, para mí, en mi propio caso. Las correlaciones evidentes entre la percatación y la conciencia en mi propio caso son tan detalladas y notables que debe haber algo más que una mera regularidad aleatoria. Debe haber alguna ley subyacente. La única pregunta es *¿qué* ley? Esta ley debe implicar que, en mi propio caso, la percatación siempre está acompañada por la conciencia y viceversa, y además que las estructuras de las dos están en correspondencia. Los principios de coherencia que formulé harán el trabajo. ¿Podría bastar también algún otro principio diferente?

Es muy plausible que algún tipo de percatación sea *necesaria* para la conciencia. Ciertamente, todas las instancias de conciencia que conozco están acompañadas por la percatación. Parece haber pocas razones para creer que existan instancias de conciencia *sin* los procesos funcionales acompañantes. Si existen, no tenemos ninguna evidencia de ellos, ni siquiera evidencia indirecta, y, en principio, no podríamos tenerla. Por consiguiente, es razonable suponer, basados en la parsimonia, que siempre que hay conciencia, hay percatación. Si nos equivocamos en esto —si por ejemplo un electrón estático tiene la rica vida consciente de un Proust— entonces seguramente nunca lo sabremos.

La cuestión de la *suficiencia* de la percatación es más difícil. Dada la necesidad de la percatación, cualquier candidato de ley subyacente tendrá la forma "Percatación más algo da origen a la conciencia". Al menos, cualquier ley subyacente debe *implicar* un principio de esta forma para poder explicar las regularidades en mi propio caso. La pregunta remanente, entonces, es: ¿Qué es el algo extra, o no se requiere nada extra?

Llamemos al ingrediente extra hipotético el *factor X*. O soy consciente en virtud de la percatación solamente, o soy consciente en virtud de la percatación y el factor X. El factor X podría ser cualquier propiedad, siempre que yo la posea en este momento y, preferible-

mente, a lo largo de mi vida. Quizás el factor X sea una cuestión de nacionalidad, y la percatación da origen a la conciencia sólo en el caso de los australianos. Quizá sea una cuestión de localización, y la percatación da origen a la conciencia únicamente dentro de un radio de cincuenta millones de kilómetros de una estrella. Quizá sea una cuestión de *identidad*, y la percatación da origen a la conciencia sólo en David Chalmers.

Todas estas leyes serían compatibles con mi evidencia y explicarían la correlación. Entonces, ¿por qué todas parecen ser tan poco razonables? Es porque en cada uno de estos casos, el factor X parece bastante *arbitrario*. No hay ninguna razón para creer que la conciencia debería depender de esas cosas; parecen ser adornos irrelevantes. No es como si el factor X tuviese un papel en la explicación de cualquiera de los fenómenos asociados con la conciencia. La percatación podría al menos ayudarnos a explicar nuestros juicios fenoménicos; estos tienen un vínculo estrecho con la conciencia, de modo que existen algunas razones para creer que existe una conexión. En cambio, cada uno de esos factores X parece salido de la nada. ¿Por qué el universo debería ser de un modo tal que la percatación dé origen a la conciencia en una persona, y por qué sólo en una persona? Sería un modo extraño y arbitrario de existencia para un mundo.

Algo parecido ocurre con los factores X "más plausibles" que alguien podría formular seriamente. Un candidato natural para un factor X de este tipo podría basarse en la biología celular o incluso en la neurofisiología humana. Algunas personas supusieron que la conciencia está limitada a seres con el tipo correcto de composición biológica. De un modo similar, algunos sugirieron que la conciencia surge de la organización funcional sólo cuando esa organización no está implementada de un modo "encabezado por homunculi", como en el ejemplo de la nación china. Pero, factores X como estos son igualmente arbitrarios. Sólo complican las leyes sin ninguna compensación agregada. ¿Por qué debería el mundo estar construido de forma que la percatación dé origen a la conciencia sólo en seres con una biología particular, o tal que los homunculi internos estén descartados? Las hipótesis parecen barrocas; parecen tener distracciones ajenas incorporadas.

¿Por qué podría alguien creer en un factor X? Creo que estas creencias surgen por una razón natural pero engañosa. Existe una intuición básica de que la conciencia es algo que va más allá de la organización funcional. Esta es una intuición que, por supuesto, comparto: la conciencia es un hecho ulterior, para el cual ninguna organización funcional es lógicamente suficiente. También existe una tendencia natural a creer que todo es físico y que la conciencia debe

ser explicable físicamente de un modo u otro. Frente a estas dos presiones, existe una reacción natural: debemos agregar algo extra y ese algo debe ser físico. La biología humana es un candidato natural para ese ingrediente extra. De esta manera, podría pensarse que hemos cruzado la brecha entre la organización funcional y la biología humana.

Pero esto está bastante descaminado. Agregar la biología al cuadro no ayudó en absoluto al problema original. La brecha es tan grande como siempre: la conciencia parece ser algo que va más allá de la biología también. Como argumenté antes, *ningún* hecho físico basta para explicar la conciencia. El factor X no puede hacer ningún trabajo para nosotros; buscamos una solución a nuestro problema en el lugar equivocado. La dificultad era, en primer lugar, el supuesto del materialismo. Una vez que aceptamos que el materialismo es falso, se hace evidente que la búsqueda de un factor X físico es irrelevante; en cambio, debemos buscar un "factor Y", algo *adicional* a los hechos físicos que nos ayude a explicar la conciencia. Encontramos un factor Y de este tipo en la postulación de leyes psicofísicas irreducibles. Una vez que las incorporamos a nuestro marco conceptual, la intuición de que la conciencia es un hecho ulterior se preserva y el problema se elimina.

El deseo de un factor X físico es un vestigio del intento de tener la torta materialista y comerse la conciencia también. Una vez que reconocemos que la conciencia es un hecho no físico ulterior y que existen leyes psicofísicas independientes, el factor X se vuelve relativamente redundante. Pedir una conexión psicofísica independiente *y* un factor X es pedir dos regalos cuando en realidad sólo necesitamos uno.

El factor X, por lo tanto, no tiene ningún papel explicativo en una teoría de la conciencia y sólo complica las cosas. Cualquier factor de este tipo hace que las leyes fundamentales sean más complejas de lo necesario. Dada la simplicidad de la imagen en la cual la percatación da origen a la conciencia, un universo en el cual la conciencia depende de un factor X separado comienza a parecer un lugar no razonable. También podríamos tener una cláusula en las leyes de Newton que diga que toda acción tiene una reacción igual y opuesta a menos que los objetos involucrados estén hechos de oro. Los principios de simplicidad dictan que la mejor hipótesis es que no se requiera ningún factor X y que la percatación dé origen a la conciencia sin salvedades.

Algunas personas todavía se sentirán inseguras respecto de la conclusión funcionalista a la que llegué, aun cuando esta sea una

versión dualista del funcionalismo. Es verdad que el argumento a partir de los factores X es algo tentativo y se basa fuertemente en supuestos de simplicidad. En el próximo capítulo enunciaré argumentos más concretos para la misma conclusión, utilizando experimentos mentales para defender la tesis de que una réplica funcional de un ser consciente tendrá precisamente la misma clase de experiencias conscientes. Pero, por ahora, observemos que estas consideraciones representan al menos un fuerte caso *prima facie* para este tipo de funcionalismo.

Vale la pena tomarse un momento para comprender el marco epistemológico general. Lo que tenemos aquí es, esencialmente, una inferencia de la mejor explicación. Advertimos regularidades notables entre la conciencia y la percatación en nuestro propio caso y postulamos la ley subyacente más simple posible. Este es el mismo tipo de razonamiento que se emplea cuando se formulan teorías físicas e incluso cuando se combaten las hipótesis escépticas acerca de la causalidad y acerca del mundo externo. En todos estos casos, el supuesto subyacente es que el mundo es un lugar simple y razonable. Si este supuesto falla, cualquier cosa vale. Adoptando el supuesto, las cosas caen en su lugar.

También, parece que esta es una solución del problema de las otras mentes tan buena como podremos llegar a obtener. Advertimos regularidades entre la experiencia y los estados físicos o funcionales en nuestro propio caso, postulamos leyes subyacentes simples y homogéneas para explicarlas, y utilizamos esas leyes para inferir la existencia de la conciencia en otros. Este podría o no ser el razonamiento que implícitamente empleamos al creer que otros son conscientes pero, de cualquier forma, parece proporcionar una justificación razonable de nuestras creencias.

Es interesante especular acerca de qué implican exactamente nuestros principios de coherencia para que exista la conciencia fuera de la especie humana y, en particular, en organismos mucho más simples. La cuestión no es clara, ya que nuestra noción de percatación sólo está claramente definida para casos que se aproximan a la complejidad humana. Parece razonable decir que un perro posee percatación y que un ratón también (quizá no tengan autopercatación, pero esa es una cuestión diferente). Por ejemplo, parece razonable decir que un perro se percata de una boca de incendio* en el sentido

* En algunos países suelen verse en las aceras unas bocas de incendio muy llamativas, de color rojo, a las que los perros se acercan habitualmente para orinar. [T.]

básico del término “percatarse”. Los sistemas de control del perro tienen acceso a la información acerca de aquella, y pueden usarla para controlar apropiadamente la conducta. Por el principio de coherencia, parece probable que el perro *experimente* la boca de incendio de un modo no muy diferente de nuestra experiencia visual del mundo. Esto concuerda con el sentido común; todo lo que hago aquí es hacer un poco más explícito el razonamiento de sentido común.

Lo mismo probablemente sea verdad para los ratones e incluso para las moscas. Las moscas tienen algún acceso perceptual limitado a la información ambiental, y puede suponerse que su contenido perceptual impregna sus sistemas cognitivos y está disponible para la conducta directa. Parece razonable suponer que esto sea una forma de percatación y que, por el principio de coherencia, exista algún tipo de experiencia acompañante. En este lugar las cosas comienzan a ponerse difíciles. Es tentador extender la coherencia aun más abajo en la escala de procesamiento de información; pero, tarde o temprano, la noción de “percatación” se agota y ya no puede hacer más trabajo explicativo, por su indeterminación. Por ahora, no continuaré especulando sobre esta cuestión, pero volveré a ella más adelante.

# 7

# Qualia ausentes, qualia desvanecientes, qualia danzantes

## 1. El principio de la invariancia organizacional

Si la conciencia surge de lo físico, ¿en virtud de qué clase de propiedades físicas surge? Es de suponer que estas serán propiedades que los cerebros pueden instanciar, pero no es obvio precisamente cuáles son las propiedades correctas. Algunos sugirieron propiedades bioquímicas; otros sugirieron propiedades cuánticas; muchos propusieron la incertidumbre. Una propuesta natural es que la conciencia surge en virtud de la *organización funcional* del cerebro. Según esta perspectiva, el sustrato químico y ciertamente el sustrato cuántico del cerebro son irrelevantes para la producción de la conciencia. Lo que cuenta es la organización causal abstracta del cerebro, una organización que podría realizarse en muchos sustratos físicos diferentes.

La organización funcional puede comprenderse mejor como el *patrón abstracto de interacción causal* entre las diversas partes de un sistema y, quizás, entre esas partes y las entradas y salidas externas. Una organización funcional se determina especificando 1) un número de componentes abstractos, 2) para cada componente, un número de estados posibles diferentes, y 3) un sistema de relaciones de dependencia que especifican cómo el estado de cada componente depende de los estados previos de todos los componentes y de las entradas en el sistema, y cómo las salidas del sistema dependen de los estados previos de los componentes. Más allá de especificar su número y sus relaciones de dependencia, la naturaleza de los componentes y los estados queda inespecificado.

Un sistema físico *realiza* una organización funcional dada cuando el sistema puede dividirse en un número apropiado de componentes físicos, cada uno con el número apropiado de estados, tal que las relaciones de dependencia causal entre los componentes del sistema, las entradas y las salidas reflejan precisamente las relaciones de dependencia definidas en la especificación de la organización funcional. (Una concepción más formal según este enfoque se ofrece en Chalmers, 1994a, 1994b y se resume en el capítulo 9, pero una comprensión informal será suficiente por ahora.)

Una organización funcional dada puede realizarse mediante diversos sistemas físicos. Por ejemplo, la organización realizada por el cerebro en el nivel neuronal podría, en principio, ser realizada por un sistema físico. Una descripción de la organización funcional del cerebro hace abstracción de la naturaleza física de las partes involucradas y del modo como se implementan las conexiones causales. Todo lo que importa es la existencia de las partes y las relaciones de dependencia entre sus estados.

Un sistema físico tiene una organización funcional en muchos niveles diferentes, dependiendo de cuán finamente individualicemos sus partes y de cuán finamente dividamos los estados de esas partes. En un nivel muy tosco, por ejemplo, es probable que pueda considerarse que los dos hemisferios del cerebro realizan una simple organización de dos componentes si elegimos estados interdependientes apropiados de los hemisferios. En general, sin embargo, es más útil considerar los sistemas cognitivos en un nivel más fino. Si estamos interesados en la cognición, usualmente nos concentraremos en un nivel lo suficientemente fino como para determinar las capacidades conductuales asociadas con el cerebro, donde la conducta se individualiza en algún nivel apropiado de precisión. La organización en un nivel muy poco detallado (por ejemplo, la organización de dos componentes antes mencionados) será insuficiente para determinar las capacidades conductuales, ya que los mecanismos que impulsan la conducta se perderán entre las fisuras de esa descripción; un sistema simple podría compartir la organización sin compartir la conducta. En un nivel lo suficientemente fino, sin embargo —quizás en el nivel neuronal—, la organización funcional determinará las capacidades conductuales. Aunque nuestras neuronas fueran reemplazadas por chips de silicio, siempre que esos chips posean estados con el mismo patrón de interacciones causales que encontramos en las neuronas, el sistema producirá la misma conducta.

En lo que sigue, el tipo apropiado de organización funcional de un sistema estará siempre en un nivel lo suficientemente fino como para determinar sus capacidades conductuales. Llamemos a esta

organización una organización funcional de *grano fino*. Para propósitos de ilustración, usualmente me concentraré en el nivel neuronal de organización en el cerebro, aunque podría bastar un nivel superior, y no es imposible que pudiese requerirse uno inferior. De cualquier manera, los argumentos pueden generalizarse. Para los propósitos de lo que sigue, necesitamos también estipular que para que dos sistemas compartan su organización funcional, deben estar en estados correspondientes en los momentos apropiados; aunque podría suponerse que mi gemelo durmiente comparte mi organización en un sentido amplio, no será así en el sentido estricto requerido más abajo. Cuando dos sistemas comparten su organización funcional en este sentido estricto, diré que son *isomorfos funcionales*.

Afirmo que la experiencia consciente surge de una organización funcional de grano fino. Más específicamente, argumentaré en favor de un *principio de invariancia organizacional*, que sostiene que dado cualquier sistema que tenga experiencias conscientes, cualquier otro sistema que tenga la misma organización funcional de grano fino tendrá experiencias cualitativamente idénticas. Según este principio, la conciencia es un invariante organizacional: una propiedad que permanece constante sobre todos los isomorfos funcionales de un sistema dado. No importa si la organización se realiza en chips de silicio, en la población de China o en latas de cerveza y pelotitas de ping-pong. En tanto la organización funcional sea la correcta, se producirá la experiencia consciente.

Esta tesis ha sido frecuentemente asociada con una perspectiva funcionalista reductiva acerca de la conciencia, como el enfoque de que todo lo que significa *ser* consciente es estar en el estado funcional apropiado. Desde este punto de vista el principio de invariancia se deduciría naturalmente, pero también puede sostenerse en forma independiente. Así como podemos creer que la conciencia surge de un sistema físico pero no es un estado físico, podemos creer que la conciencia surge de la organización funcional pero no es un estado funcional. El enfoque que propuse tiene esa forma; podríamos llamarlo *funcionalismo no reductivo*. Se lo podría considerar un modo de combinar el funcionalismo y el dualismo de propiedades.

En lo que sigue no me ocuparé especialmente de los aspectos no reductivos de mi enfoque; me dedicaré principalmente a argumentar en favor del principio de invariancia. Mis argumentos podrían ser aceptados incluso por los funcionalistas reductivos. Aunque los argumentos no establecen la conclusión reductiva completa, de todos modos se puede considerar que apoyan esa posición en contra de otros enfoques reductivos, como uno en el que la conciencia se equipara con una propiedad bioquímica. Por supuesto, creo que todos los enfoques

reductivos finalmente fracasan, pero la siguiente exposición será básicamente independiente de esa cuestión.

Ya argumenté en favor de una versión del principio de invariancia mediante el argumento del "factor X" en el capítulo 6. En este capítulo, sin embargo, utilizaré experimentos mentales para defender el principio de un modo mucho más directo.

### Qualia ausentes y qualia invertidos

El principio de invariancia está lejos de ser universalmente aceptado. Muchos dualistas y materialistas argumentaron en su contra. Muchos sostuvieron que para que un sistema sea consciente, debe tener el tipo correcto de conformación bioquímica; si esto es así, un robot metálico o un ordenador basado en silicio nunca podrán tener experiencias, sin importar cuál sea su organización causal. Otros aceptaron que un robot o un ordenador podrían ser conscientes si estuvieses organizados apropiadamente, pero sostuvieron que no obstante tendrían experiencias bastante diferentes de las nuestras.

En correspondencia con estos dos enfoques, existen en general dos clases de argumentos en contra del principio de invariancia. La primera clase comprende los argumentos a partir de los *qualia ausentes*. En estos argumentos, se describe una realización particularmente excéntrica de una organización funcional dada, en un sistema tan estrafalario que es natural suponer que las cualidades (qualia) de la experiencia consciente deben estar *ausentes*. Un ejemplo popular de Block (1978) es un caso en el cual nuestra organización se realiza en la población de un país (como en el capítulo 3). Seguramente, se argumenta, *eso* no podría dar origen a la experiencia consciente. Si esto es así, entonces la conciencia no puede surgir de la organización funcional.

La segunda clase de argumentos incluye aquellos a partir de los *qualia invertidos*, o el *espectro invertido*. Según estos argumentos, si nuestra organización funcional se realizase en un sustrato físico distinto, un sistema podría todavía tener experiencias, pero serían de un tipo diferente. Cuando nosotros tenemos experiencias de rojo, el sistema podría tener experiencias de azul, etc. Estos argumentos suelen hacerse por medio de escenarios complejos que recurren a la cirugía cerebral, en los que nos despertamos una mañana viendo azul en lugar de rojo aunque nuestra organización funcional no se haya modificado.

Muchos de los que argumentan acerca de la posibilidad de los qualia ausentes e invertidos sólo lo hacen en cuanto a su posibilidad lógica; esto es todo lo que se requiere para refutar una forma reductiva

del funcionalismo. Yo mismo utilicé argumentos de este tipo en el capítulo 3. Estos proponentes no están sujetos a los contraargumentos de este capítulo. Lo que está en cuestión aquí es una forma más débil de funcionalismo, una que no depende de cuestiones de posibilidad lógica.

La cuestión clave en este capítulo es si los qualia ausentes o invertidos son *natural* (o *empíricamente*) posibles. Es lógicamente posible que un plato pueda flotar hacia arriba cuando lo liberamos en el vacío sobre la superficie de un planeta pero, no obstante, es naturalmente imposible. Las leyes de la naturaleza lo prohíben. De un modo similar, establecer la posibilidad lógica de los qualia ausentes y los qualia invertidos no es suficiente para establecer su posibilidad natural. El principio de invariancia sostiene que la organización funcional determina la experiencia consciente mediante algún vínculo legaliforme en el mundo real; aquí, las cuestiones de posibilidad lógica son irrelevantes. En este capítulo, cada vez que utilizamos el término "posibilidad" sin modificador, lo que queremos significar es posibilidad natural.

En lo que sigue, analizaré los argumentos formulados en favor de la posibilidad natural de los qualia ausentes e invertidos, y luego ofreceré argumentos detallados *en contra* de esas posibilidades. Esos argumentos involucrarán de un modo crucial experimentos mentales. En contra de la posibilidad de los qualia ausentes, ofreceré un experimento mental que concierne a *qualia desvanecientes.* En contra de la posibilidad de los qualia invertidos, ofreceré un experimento mental que concierne a *qualia danzantes.*

Como siempre, estos argumentos a partir de experimentos mentales sólo son argumentos de plausibilidad, pero creo que tienen una fuerza considerable. Para mantener la posibilidad natural de los qualia ausentes e invertidos frente a estos experimentos mentales se requiere aceptar algunas tesis poco razonables acerca de la naturaleza de la experiencia consciente y, en particular, acerca de la relación entre la conciencia y la cognición. Dados ciertos supuestos naturales acerca de esta relación, el principio de invariancia surge como la hipótesis más plausible.

Quizá sea útil considerar que estos experimentos mentales tienen un papel análogo al del experimento mental del "gato de Schrödinger" en la interpretación de la mecánica cuántica. El experimento mental de Schrödinger no produce un veredicto decisivo en favor de una interpretación u otra, pero pone de relieve diversas plausibilidades e inverosimilitudes en las interpretaciones, y es algo que todas las interpretaciones deben finalmente poder captar. De un modo similar, cualquier teoría de la conciencia debe finalmente llegar

a captar los escenarios de los qualia desvanecientes y danzantes, y algunas teorías los manejarán mejor que otras. De esta manera pueden clarificarse las virtudes e inconvenientes de las diversas teorías.

## 2. Qualia ausentes

Los argumentos positivos en favor de la posibilidad natural de los qualia ausentes no tuvieron la misma importancia que los argumentos en favor de los qualia invertidos, pero recibieron algunas formulaciones. La presentación más detallada de estos argumentos fue realizada por Block (1978).

Estas consideraciones casi siempre tienen la misma forma. Consisten en la exhibición de una realización de nuestra organización funcional en algún medio inusual, combinada con una apelación a la intuición. Se señala, por ejemplo, que la organización de nuestro cerebro podría ser simulada por el pueblo de China o incluso reflejarse en la economía de Bolivia. Si consiguiésemos que cada persona en China simulase una neurona (deberíamos multiplicar la población por diez o cien, pero no importa) y los equipásemos con enlaces de radio para simular conexiones sinápticas, entonces estaría definida la organización funcional. Pero, seguramente, dice el argumento, ¡este sistema barroco no será *consciente*!

Este argumento posee una cierta fuerza intuitiva. Muchas personas tienen una fuerte sensación de que un sistema como este es simplemente el tipo erróneo de cosas que podrían tener experiencias conscientes. Una "mente grupal" de esta clase parecería el tema de un cuento de ciencia ficción, más que el tipo de cosas que podrían realmente existir. Pero la fuerza de esto es *sólo* intuitiva. No configura un argumento devastador. Muchos señalaron[1] que aunque puede ser intuitivamente inverosímil que un sistema de esta clase pueda dar origen a la experiencia, ¡también es intuitivamente inverosímil que un *cerebro* pueda dar origen a la experiencia! ¿Quién habría pensado que este trozo de materia gris sería el tipo de cosa que podría producir experiencias subjetivas vívidas? Y sin embargo lo hace. Por supuesto, esto no *muestra* que la población de una nación pueda producir una mente, pero es un fuerte contraargumento al argumento intuitivo de que no lo hará.

Por supuesto, no *veríamos* ninguna experiencia consciente en un sistema de esta clase. Pero esto no es nuevo; no vemos las experiencias conscientes de nadie. Podría parecer que no hay "espacio" para la experiencia consciente en un sistema de este tipo pero, nuevamente, lo mismo parece ser verdad del cerebro. Tercero, podríamos *explicar*

el funcionamiento del sistema sin invocar la experiencia consciente pero, nuevamente, esto es familiar en el caso estándar. Una vez que absorbemos la verdadera fuerza de la no superveniencia lógica, comienza a parecernos que el hecho de que la población de un país pueda dar origen a la experiencia consciente no es más sorprendente que un cerebro pueda hacerlo.[2]

Algunos objetaron el principio de invariancia sobre la base de que la organización funcional podría surgir por azar tanto en la economía boliviana como en un cubo (Hinckfuss, citado en Lycan, 1987, p. 32). Pero esto sólo podría ocurrir por la coincidencia más escandalosa.[3] El sistema debería tener más de mil millones de partes, cada una con un cierto número de estados (digamos, diez cada uno). Entre estos estados debería haber un sistema vasto e intrincado de exactamente las conexiones causales correctas, de modo que dado *este* patrón de estados, resultará ese otro patrón de estados, etc. Para realizar la organización funcional en cuestión, estos condicionales no pueden ser meras regularidades (en las que este patrón de estados es seguido por ese otro patrón de estados en esta ocasión); deben ser conexiones fiables que sustenten contrafácticos, de modo que este patrón de estados será seguido por ese otro patrón de estados cada vez que surja.[4]

No es difícil ver que aproximadamente $10^{10^9}$ condicionales de este tipo serán necesarios en un sistema para que realice la organización funcional apropiada, si suponemos una división en mil millones de partes. La probabilidad de que estos condicionales puedan ser satisfechos por un sistema arbitrario bajo una división dada en partes y estados es del orden de 1 en $(10^{10^9})^{10^{10^9}}$ (en realidad, es mucho menor, ya que el requerimiento de que cada condicional sea fiable reduce mucho más las posibilidades de que sea satisfecho).[5] Aun dada la libertad que tenemos para dividir un sistema en partes, es extraordinariamente improbable que una organización de este tipo sea realizada por un sistema arbitrario o, ciertamente, por *cualquier* sistema que no haya sido formado por los mecanismos altamente no arbitrarios de la selección natural.

Una vez que advertimos cuán estrechamente la especificación de una organización funcional constriñe la estructura de un sistema, se hace más plausible que hasta la población de China pueda realizar la experiencia consciente si está organizada de forma apropiada. Si tomamos nuestra imagen de la población, la aceleramos en un factor de un millón aproximadamente, y la encogemos a un área del tamaño de una cabeza, nos queda algo que se parece mucho a un cerebro, excepto que tiene homunculi —pequeñas personas— en donde un cerebro tendría neuronas. A primera vista, no hay muchas razones

para suponer que las neuronas deberían hacer un mejor trabajo que los homunculi para realizar la experiencia.

Por supuesto, como Block señala, *sabemos* que las neuronas pueden hacer el trabajo, mientras que no lo sabemos acerca de los homunculi. La cuestión, por lo tanto, permanece abierta. El punto importante es que este tipo de argumento sólo representa una evidencia muy débil de que los qualia ausentes son naturalmente imposibles. Se requiere un argumento más convincente para decidir la cuestión de un modo u otro. *Quizá* sea correcto decir, como Block lo hace, que nuestras intuiciones colocan la carga de la prueba sobre quien sostenga que los qualia son organizacionalmente invariantes, aunque esto no me resulta del todo evidente. De cualquier forma, aceptaré esa carga en lo que sigue.

Un argumento independiente que a veces se formula en favor de la posibilidad natural de los qualia ausentes surge del fenómeno de la ceguera visual. Se argumenta que los pacientes con ceguera visual son funcionalmente similares a nosotros en los modos apropiados —pueden discriminar, comunicar contenidos, etc.— pero carecen de la experiencia visual. Por lo tanto, la organización funcional del procesamiento visual no determina la presencia o ausencia de la experiencia.

Hemos visto en el capítulo 6 que existe una *diferencia* significativa entre el procesamiento en sujetos normales y en aquellos que padecen de ceguera visual, sin embargo. Estos sujetos carecen del tipo usual de acceso directo a la información visual. En los casos en los que la información es accesible, el acceso es indirecto y la información no está disponible para el control de la conducta del modo usual. Es precisamente debido a la diferencia en la organización de su procesamiento, tal como se manifiesta en su conducta, que advertimos algo inusual en primer lugar y nos vemos llevados a postular la ausencia de experiencia. Estos casos, por lo tanto, no proporcionan evidencia en contra del principio de invariancia.

## 3. Qualia desvanecientes

Mi argumento positivo en contra de la posibilidad de los qualia ausentes se basará en un experimento mental que involucra el reemplazo gradual de partes de un cerebro por, quizá, chips de silicio. Los experimentos mentales de este tipo han sido una respuesta popular a los argumentos de los qualia ausentes en la tradición folclórica de la inteligencia artificial y a veces en trabajos publicados. Pylyshyn (1980) examina el escenario del reemplazo gradual, aunque sin un argumento sistemático que acompañe ese examen. Argumen-

tos similares al que formularé fueron planteados por Savitt (1982) y Cuda (1985), aunque ellos desarrollan los argumentos de modos diferentes y extraen moralejas levemente distintas a partir del escenario.[6]

Este argumento de los "qualia desvanecientes" no será mi argumento más fuerte ni principal en contra de la posibilidad de los qualia ausentes; ese papel lo tendrá el argumento de los "qualia danzantes", que desarrollaré en el apartado 5, y que también representa un argumento en contra de la posibilidad de los qualia invertidos. Sin embargo, el argumento de los qualia desvanecientes es fuerte en sí mismo y proporciona una buena motivación y telón de fondo para el más poderoso segundo argumento.

El presente argumento toma la forma de una *reductio ad absurdum*. Supóngase que los qualia ausentes son naturalmente posibles. Entonces podría haber un sistema con la misma organización funcional que un sistema consciente (digamos, yo mismo), pero que carece totalmente de experiencias conscientes. Sin pérdida de generalidad, supongamos que esto se debe a que el sistema está hecho de chips de silicio en lugar de neuronas. Más adelante mostraré cómo el argumento puede extenderse a otros tipos de isomorfos. Llamemos a este isomorfo funcional Robot. Los patrones causales en el sistema cognitivo de Robot son los mismos que los míos, pero él tiene la conciencia de un zombi.

Dada esta situación, podemos construir una serie de casos intermedios entre mi persona y Robot tal que en cada paso únicamente haya un cambio muy pequeño y tal que la organización funcional se preserve en todo momento. Podemos imaginar, por ejemplo, que reemplazamos un cierto número de mis neuronas por chips de silicio. En el primer paso, sólo reemplazamos una neurona. Su reemplazo es un chip de silicio que realiza precisamente la misma función local que la neurona. Allí donde esta se conecta con otras neuronas, el chip se conecta con esas mismas neuronas. Cuando el estado de la neurona es sensible a entradas eléctricas y señales químicas, el chip de silicio es sensible a las mismas entradas. Podríamos imaginar que el chip está equipado con pequeños transductores que aceptan las señales eléctricas y iones químicos y que transmiten una señal digital al resto del chip. Cuando la neurona produce salidas eléctricas y químicas, el chip hace lo mismo (podemos imaginar que está equipado con diminutos efectores que producen salidas eléctricas y químicas según el estado interno del chip). Es importante que los estados internos del chip sean tales que la función de entrada/salida del chip sea precisamente la misma que la de la neurona. No importa cómo el chip lo haga —podría utilizar un cuadro de doble entrada que asocia

cada entrada con la salida apropiada o quizá realice una computación que simula los procesos dentro de una neurona— en tanto produzca las dependencias de entrada-salida correctas. Si las interacciones locales son correctas, entonces el reemplazo no hará ninguna diferencia para la función general del sistema.

En el segundo paso, reemplazamos dos neuronas por chips de silicio. Será más fácil suponer que son neuronas vecinas. De este modo, una vez que ambas fueron reemplazadas podemos eliminar los torpes transductores y efectores que median la conexión entre los dos chips. Podemos reemplazarlos por cualquier tipo de conexión que nos guste, siempre que sea sensible al estado interno del primer chip y afecte el estado interno del segundo chip del modo apropiado (puede haber una conexión en cada dirección, por supuesto). Aquí nos aseguramos de que la conexión sea una copia de la conexión correspondiente en Robot; quizás esta sea una señal electrónica de algún tipo.

Los pasos posteriores proceden del modo obvio. En cada paso siguiente un grupo cada vez mayor de neuronas vecinas se reemplaza por chips de silicio. Dentro de este grupo de chips, el sustrato bioquímico ha sido eliminado totalmente. Los mecanismos bioquímicos están presentes sólo en el resto del sistema y en la conexión entre los chips en la frontera del grupo y las neuronas vecinas. En el caso final, toda neurona del sistema ha sido reemplazada por un chip y no existen mecanismos bioquímicos que desempeñen un papel esencial. (Hago abstracción aquí de cuestiones de detalle como, por ejemplo, la de si las células gliales desempeñan un papel no trivial; si esto es así, serán componentes de la organización funcional apropiada y también serán reemplazadas.)

Podemos imaginar que, en todo momento, el sistema interno está conectado a un cuerpo, es sensible a las entradas sensoriales y produce movimientos motores de un modo apropiado, por medio de transductores y efectores. Cada sistema en la secuencia será funcionalmente isomórfico conmigo en un grano suficientemente fino como para compartir mis disposiciones conductuales. Pero, mientras que el sistema en un extremo del espectro soy yo, el sistema en el otro extremo es esencialmente una copia de Robot.

Para fijar ideas, supóngase que como primer sistema estoy teniendo ricas experiencias conscientes. Quizás esté en un partido de baloncesto, rodeado de aficionados que gritan, con todo tipo de vestimentas coloridas a mi alrededor, oliendo el delicioso aroma de la comida chatarra, quizá padeciendo un dolor palpitante de cabeza, etc. Concentrémonos, en particular, en las experiencias de rojo brillante y amarillo que tengo al mirar los uniformes de los jugadores. El

sistema final, Robot, está en la misma situación, procesa las mismas entradas y produce conductas similares pero, según la hipótesis, no experimenta nada en absoluto.

Entre mi persona y Robot, habrá muchos casos intermedios. Pregunta: *¿Cómo es ser como ellos?* ¿Qué experimentan, si experimentan algo? A medida que nos movemos a lo largo del espectro, ¿cómo varía la experiencia consciente? Presumiblemente los casos muy tempranos tienen experiencias muy similares a las mías, y los casos muy posteriores tienen poca o ninguna experiencia pero, ¿qué ocurre con los casos intermedios?

Dado que el sistema en el otro extremo del espectro (Robot) no es consciente, pareciera que una de dos cosas debe ocurrir en el camino. O 1) la conciencia se desvanece gradualmente a lo largo de la serie de casos antes de por fin desaparecer, o 2) en algún lugar en el camino, la conciencia de pronto se apaga, aunque el caso precedente tenía ricas experiencias conscientes. Llamemos a la primera posibilidad *qualia gradualmente desvanecientes* y a la segunda *qualia repentinamente desvanecientes.*

No es difícil desechar la posibilidad de los qualia repentinamente desvanecientes. Según esta hipótesis, el reemplazo de una sola neurona (mientras todo el resto permanece constante) podría ser responsable del desvanecimiento de un campo entero de experiencia consciente. Esto parece extremadamente inverosímil, si no totalmente extravagante. Si esto fuera posible, existirían abruptas discontinuidades en las leyes de la naturaleza, a diferencia de lo que encontramos en cualquier otro lado. Cualquier punto específico en el que los qualia repentinamente desaparecieran (¿50% de las neuronas?, ¿25%?) sería totalmente arbitrario. Podemos incluso imaginarnos realizar el experimento mental en un grano más fino dentro de la neurona, de modo que, finalmente, el reemplazo de unas pocas moléculas causaría que todo un campo de experiencia desaparezca (si esto no es así, revertimos al escenario de los qualia gradualmente desvanecientes). Como siempre en estas cuestiones, la hipótesis no puede ser refutada, pero su plausibilidad es escasa.

(Podríamos argumentar que existen situaciones en la dinámica no lineal en las cuales una magnitud depende sensiblemente de otra, donde grandes cambios en la primera surgen de pequeños cambios en la segunda. Pero, en estos casos la dependencia es, no obstante, continua, de modo que habrá casos intermedios en los cuales la magnitud dependiente toma valores intermedios; la analogía, por lo tanto, lleva a los qualia gradualmente desvanecientes. De cualquier forma, la dependencia en estos casos surge, por lo general, de los efectos compuestos de un número de dependencias graduales más

básicas. En todas las leyes fundamentales conocidas hasta la fecha, la dependencia de una magnitud sobre otra magnitud continua es continua de este modo, y no hay forma de componer continuidad en discontinuidad. Los qualia repentinamente desvanecientes, en contraste con la dinámica no lineal, requeriría discontinuidades primitivas en las leyes fundamentales.)

Si se descartan los qualia repentinamente desvanecientes, nos quedan los qualia gradualmente desvanecientes. Para situarnos en este escenario, considérese un sistema a mitad de camino a lo largo del espectro entre mi persona y Robot, luego de que la conciencia se degradó considerablemente pero antes de que haya desaparecido del todo. Llamemos a este sistema José. ¿Cómo es ser José? José, por supuesto, es funcionalmente isomórfico a mí mismo. El *dice* todas las mismas cosas acerca de sus experiencias que yo digo acerca de las mías. En el juego de baloncesto, habla acerca de los uniformes de rojo y amarillo brillante de los jugadores.

Según la hipótesis, sin embargo, José no está teniendo experiencias de rojo y amarillo brillantes en absoluto. En cambio, quizá sólo experimente un tenue rosa y un marrón oscuro. Quizás tenga la más débil de las experiencias de rojo y amarillo. Tal vez sus experiencias se oscurecieron casi hasta el negro. Hay diversos modos concebibles en los que las experiencias de rojo podrían gradualmente transmutarse en ninguna experiencia y es posible que haya aun más modos que no podemos concebir. Pero, puede suponerse que en cada una de estas maneras las experiencias deben dejar de ser *brillantes* antes de desaparecer (si no nos encontraríamos con el problema de los qualia repentinamente desvanecientes). De modo similar, podemos suponer que existe un punto en el que distinciones sutiles en mi experiencia ya no están presentes en la experiencia de un sistema intermedio; si suponemos que todas las distinciones en mi experiencia están presentes hasta el momento en el que todas desaparecen simultáneamente, volvemos a encontrar otra versión de los qualia repentinamente desvanecientes.

Para especificar, entonces, imaginemos que José ve un rosa lavado cuando yo veo un rojo brillante y que muchas distinciones entre los matices de mi experiencia ya no están presentes en la suya. Cuando yo tengo experiencias de ruidos fuertes, quizá José experimente sólo un distante fragor. No todo es tan malo para José: cuando yo tengo un dolor palpitante de cabeza, él sólo tiene una tenue puntada.

La característica crucial aquí es que José está sistemáticamente *equivocado* acerca de todo lo que experimenta. El *dice* que tiene experiencias de rojo y amarillo brillantes, pero sólo experimenta un rosa tenue.[7] Si usted le pregunta, él afirmará estar experimentando

todo tipo de matices sutilmente diferentes de rojo pero, de hecho, muchos de estos son bastante homogéneos en su experiencia. Puede incluso quejarse del ruido, cuando su experiencia auditiva está realmente muy amortiguada. Pero incluso en una interpretación funcional de las creencias, José *creerá* que tiene todas esas experiencias complejas de las que de hecho carece. En síntesis, José está completamente fuera de contacto con su experiencia consciente y es incapaz de ponerse en contacto con ella.

Esto parece bastante inverosímil. Tenemos aquí un ser cuyos procesos racionales funcionan y que es de hecho *consciente*, pero que está completamente equivocado acerca de sus experiencias conscientes. Quizás en el caso extremo, cuando todo es oscuro adentro, pueda ser razonable suponer que un sistema podría estar tan descarriado en sus afirmaciones y juicios; después de todo, en cierto sentido, no hay nadie allí adentro para equivocarse. Pero, en el caso intermedio, esto es mucho menos plausible. En cada caso con el que estamos familiarizados, los seres conscientes son por lo general capaces de formar juicios precisos acerca de su experiencia, en ausencia de distracción e irracionalidad. Para un ser sensible y racional que no padece ninguna patología funcional, estar tan sistemáticamente fuera de contacto con sus experiencias significaría suponer una fuerte disociación entre la conciencia y la cognición. Tenemos pocas razones para creer que la conciencia sea un fenómeno con una conducta tan anómala de este tipo y buenas razones para creer lo contrario.

Con seguridad, los qualia gradualmente desvanecientes son *lógicamente* posibles. No hay ninguna contradicción en la descripción de un sistema que esté tan equivocado acerca de sus experiencias.[8] Pero la posibilidad lógica y la posibilidad natural son cosas diferentes. No tenemos ninguna razón para creer que este tipo de casos pueda ocurrir en la práctica y todas las razones para creer lo contrario. Uno de los hechos empíricos más sobresalientes acerca de la conciencia parece ser que cuando un ser consciente con la sofisticación conceptual apropiada tiene experiencias, es capaz de formar juicios acerca de esas experiencias. Quizás existan casos inusuales en los que los procesos racionales en un sistema puedan estar fuertemente deteriorados, lo que llevaría a un mal funcionamiento de los mecanismos de juicio, pero este no es un caso de ese tipo. Los procesos de José *funcionan* tan bien como los míos; por hipótesis, él es funcionalmente isomórfico a mí. Sólo ocurre que él está completamente descaminado acerca de su experiencia.

Por supuesto, existen varios casos de qualia gradualmente desvanecientes en la vida cotidiana. Piense en lo que ocurre cuando nos estamos quedando dormidos; o en una vuelta hacia atrás en la

cadena evolutiva desde las personas a los trilobites. En cada caso, cuando nos movemos a lo largo de un espectro de casos, el desvanecimiento está acompañado por un cambio correspondiente en el *funcionamiento*. Cuando me adormezco, no creo que estoy totalmente despierto y teniendo experiencias intensas (a menos, quizá, que comience a soñar, en cuyo caso es muy probable que *esté* teniendo experiencias intensas). La falta de riqueza en las experiencias del color de un perro está acompañada por una falta correspondiente de potencia discriminatoria en sus mecanismos visuales. Estos casos son bastante diferentes del que aquí consideramos, en el cual la experiencia se desvanece mientras que el funcionamiento permanece constante. Los mecanismos de José todavía pueden discriminar longitudes de ondas lumínicas sutilmente diferentes y juzga que esas discriminaciones se reflejan en su experiencia, pero tenemos razones para creer que su experiencia no refleja esas discriminaciones en absoluto.

Searle (1992) analiza un experimento mental como este, y sugiere la siguiente posibilidad:

> A medida que se implanta progresivamente el silicio en su menguante cerebro, usted encontrará que el área de su experiencia consciente se encoge, pero que esto no tiene ningún efecto sobre su conducta externa. Usted advierte, para su total sorpresa, que comienza a perder el control de su conducta externa. Descubre, por ejemplo, que cuando los médicos examinan su visión, usted les escucha decir, "Estamos sosteniendo un objeto rojo frente a usted; por favor, díganos qué ve". Usted quiere gritar, "No puedo ver nada. Estoy totalmente ciego". Pero escucha su voz diciendo de un modo que está completamente fuera de su control, "Veo un objeto rojo frente a mí". Si llevamos el experimento mental al límite, obtenemos un resultado mucho mas deprimente que la última vez. Nos imaginamos que su experiencia consciente lentamente se encoge hasta la nada, mientras que su conducta externamente observable sigue siendo la misma. (pp. 66-67)

Aquí Searle acepta la posibilidad de los qualia gradualmente desvanecientes, pero sugiere que un sistema de este tipo no necesita estar equivocado en sus creencias acerca de su experiencia. El sistema podría tener creencias verdaderas de su experiencia; sólo que estas creencias son impotentes para afectar su conducta.

Parece, sin embargo, que esta posibilidad puede descartarse. Simplemente, no hay espacio en el sistema para que se forme ninguna nueva creencia. A menos que seamos dualistas de una variedad muy fuerte, este tipo de diferencia en las creencias debe reflejarse en el

funcionamiento de un sistema, *tal vez* no en la conducta, pero al menos en algún proceso. Pero este sistema es idéntico al sistema original (yo mismo) en un grano más fino. Simplemente no hay espacio para nuevas creencias como "No puedo ver nada", para nuevos deseos como el deseo de alzar la voz u otros nuevos estados cognitivos como la sorpresa. No hay espacio para ello en las neuronas, que después de todo son idénticas a un subconjunto de las neuronas que sustentan las creencias usuales; y ¡Searle seguramente no sugiere que el reemplazo por silicio soporte las nuevas creencias! Si no ocurre ningún efecto mágico y sorprendente de interacción entre las neuronas y el silicio —uno que no se manifieste en ningún lugar del procesamiento, ya que la organización se preserva en todo momento— esas nuevas creencias no surgirán.

Una sustitución de neuronas por silicio que preserve la organización simplemente no representa una alteración tan grande como para efectuar un cambio notable en el contenido y la estructura de nuestros estados cognitivos. Un giro en la experiencia del rojo al azul es una cosa, pero un cambio en las creencias de "Buen partido de baloncesto" a "¡Oh no, me parece que estoy metido en una mala película de terror!" es de un orden de magnitud diferente. Si un cambio importante de esta clase en el contenido cognitivo no se reflejase en un cambio en la organización funcional, la cognición flotaría libre del funcionamiento interno como una mente cartesiana incorpórea. Si el contenido de los estados cognitivos superviniese a los estados físicos, sólo lo podría hacer por medio de las reglas más arbitrarias y caprichosas (si ocurre esta organización en las neuronas, entonces "¡Lindos colores!"; si ocurre en el silicio, entonces "¡Ay!").

Se deduce que la posibilidad de los qualia gradualmente desvanecientes requiere una relación extravagante entre los contenidos de las creencias y los estados físicos, o la posibilidad de seres que estén masivamente equivocados acerca de sus propias experiencias conscientes a pesar de ser totalmente racionales. Ambas hipótesis son mucho menos plausibles que la hipótesis de que los seres conscientes racionales por lo general tienen razón en sus juicios acerca de sus experiencias. Una hipótesis mucho más razonable es, por lo tanto, que cuando se reemplazan las neuronas, los qualia no se desvanecen en absoluto. Un sistema como José tendrá, en la práctica, experiencias conscientes tan ricas como las mías. Si esto es así, entonces, nuestro supuesto original era erróneo y el isomorfo original, Robot, tiene experiencias conscientes.

El argumento puede extenderse fácilmente a otros isomorfos funcionales. Para tratar con el caso en el que la población de un país implementa mi organización, podemos construir un espectro similar

de casos entre mi isomorfo de silicio y la población. Tal vez podríamos primero expandir gradualmente el sistema de silicio hasta que tenga muchos kilómetros cuadrados de extensión. También disminuimos su velocidad de procesamiento de modo que los chips reciban entradas a una velocidad manejable. Después de hacer esto, hacemos que las personas reemplacen una por vez a los chips, asegurándonos de que inician las salidas en forma apropiada en respuesta a las entradas. Eventualmente, tendremos un caso en el que toda la población está organizada como lo estaban mis neuronas, quizás, incluso, podríamos hacer que controle un cuerpo mediante enlaces de radio. En cada etapa, el sistema será funcionalmente isomórfico conmigo mismo y podemos aplicar precisamente los mismos argumentos. La experiencia consciente se preservará o se desvanecerá gradual o repentinamente. Las dos últimas hipótesis serán tan inverosímiles como antes. Podemos concluir que el sistema de la población sustentará experiencias conscientes, como el cerebro lo hace.

Podemos hacer lo mismo con cualquier sistema funcionalmente isomórfico, incluyendo aquellos que difieren en forma, tamaño, velocidad, conformación física, etc. En todos los casos, la conclusión es la misma. Si un sistema de este tipo no es consciente, entonces existe un sistema isomórfico intermedio que lo es, que tiene experiencias desvanecientes y que está completamente equivocado acerca de sus experiencias. A menos que estemos preparados para aceptar esta disociación masiva entre la conciencia y la cognición, el sistema original debe, después de todo, haber sido consciente.

Si los qualia ausentes son posibles, entonces los qualia desvanecientes también lo son. Pero argumenté antes que los qualia desvanecientes son casi imposibles. Se deduce, entonces, que los qualia ausentes son casi imposibles.

Examinaremos ahora las diversas objeciones a este argumento.

### Objeción 1: El reemplazo neuronal sería imposible en la práctica

Los que tengan inclinaciones prácticas podrían no sentirse impresionados con esta metodología de experimentos mentales. Podrían objetar que el reemplazo de neuronas por chips de silicio es materia de ciencia ficción, no materia de realidad. En particular, podrían objetar que este tipo de reemplazo sería imposible en la práctica, de modo que cualquier conclusión que pudiera obtenerse no reflejaría la realidad de la situación.

Si se supone que esto es sólo una imposibilidad *técnica*, no constituye un gran problema. Lo que está en cuestión aquí es qué

clase de experiencia tendría un sistema de ese tipo *si* existiese, sea que podamos construirlo o no. La imposibilidad natural, sin embargo, podría ser significativa. Quizás el silicio simplemente carezca de la capacidad para realizar las funciones en el cerebro que una neurona realiza, de modo que ningún chip de silicio podría estar a la altura de la tarea. No es claro que exista una razón de principio para esta objeción; ya tenemos brazos y piernas protésicas, ojos protésicos están en camino, de modo que ¿por qué no neuronas protésicas? En cualquier caso, aun si un isomorfo funcional de silicio fuera imposible (¿quizá porque la función neuronal no es computable?), el argumento en favor del principio de invariancia no se vería afectado. El principio de invariancia sólo dice que *si* existe un isomorfo funcional de un sistema consciente, *entonces* tendrá el mismo tipo de experiencias conscientes. Si los isomorfos de silicio son imposibles, el análisis de sistemas de silicio es aquí simplemente irrelevante.

Un opositor podría intentar concentrarse en los problemas de la *interfaz* silicio/neurona, en cuyo caso el sistema neuronal puro y el sistema de silicio puro serían ambos posibles, pero los sistemas intermedios podrían ponerse en cuestión. ¿Podría ocurrir que simplemente no haya suficiente espacio para los transductores y efectores en el diminuto espacio que un chip tiene disponible? Después de todo, los efectores deberían almacenar un depósito de sustancias químicas para poderlas secretar cuando fuera necesario. Pero sólo necesitamos un pequeño depósito; ¡el argumento sólo requiere el isomorfismo por unos pocos segundos! Y siempre sería posible realizar el experimento mental suponiendo una expansión del sistema. En cualquier caso, es difícil ver cómo esta cuestión sustentaría objeciones principistas profundas al principio de invariancia. Puede suponerse que habrá *algunos* sistemas entre los cuales es posible el reemplazo gradual; ¿los objetantes argumentarán que el principio de invariancia vale para esos sistemas, pero no para otros? Si esto es así, la situación parece bastante arbitraria; si no, entonces debería haber objeciones más profundas disponibles.

### Objeción 2: Algunos sistemas están masivamente equivocados acerca de su experiencia

Esta objeción observa que existen casos reales en los que los sujetos están seriamente equivocados acerca de sus experiencias. En los casos de negación de ceguera, por ejemplo, los sujetos creen que tienen experiencias visuales cuando es probable que no tengan ninguna. En estos casos, sin embargo, ya no tratamos con sistemas totalmente racionales. En sistemas cuyos mecanismos de formación

de creencias están deteriorados, cualquier cosa puede ocurrir. Sistemas de esta clase podrían creer que son Napoleón o que la luna es rosa. Mi isomorfo "desvaneciente" José, en cambio, es un sistema totalmente racional, cuyos mecanismos cognitivos funcionan tan bien como los míos. En la conversación, él parece perfectamente razonable. No podemos señalar ninguna conexión inferencial inusualmente pobre entre sus creencias, ni ningún desorden psiquiátrico sistemático que haga que sus procesos de pensamiento estén predispuestos hacia un razonamiento defectuoso. José es una persona eminentemente reflexiva y razonable, que no exhibe ninguno de los síntomas confabulatorios de aquellos que sufren negación de ceguera. Los casos entonces no son análogos. La aseveración plausible no es que ningún sistema pueda estar masivamente equivocado acerca de sus experiencias, sino que ningún sistema racional cuyos mecanismos cognitivos no están deteriorados puede estar tan equivocado. José es ciertamente un sistema racional cuyos mecanismos funcionan tan bien como los míos, así que el argumento no resulta afectado.

### Objeción 3: Los argumentos sorites son sospechosos

Algunos objetan que este argumento tiene la forma de un argumento sorites y observan que estos argumentos son, por lo general, sospechosos. Utilizando un argumento sorites, podemos "mostrar" que hasta un grano de arena es un montículo; después de todo, un millón de granos de arena forman un montículo y si quitamos un solo grano de un montículo todavía tenemos un montículo. Esta reacción, sin embargo, se basa en una lectura superficial del argumento. Los argumentos sorites por lo general obtienen su fuerza de ignorar el hecho de que alguna aparente dicotomía es, de hecho, un continuo: hay todo tipo de casos vagos entre montículos y no montículos, por ejemplo. Mi argumento, en cambio, acepta explícitamente la posibilidad de un continuo, pero sostiene que los casos intermedios son imposibles por razones independientes.

El argumento sería un sorites si tuviese la forma: Yo soy consciente; si reemplazamos una neurona en un sistema consciente por un chip de silicio este todavía será consciente; por lo tanto un sistema totalmente de silicio será consciente. Pero esta no es su forma. Es verdad que el argumento en contra de los qualia repentinamente desvanecientes se basa en la imposibilidad de una transición súbita, pero es importante notar que se opone a las transiciones repentinas *grandes*, es decir, de ricas experiencias conscientes a ninguna en absoluto. Esto es inverosímil por razones independientes de las consideraciones de los sorites.[9]

### Objeción 4: Argumentos similares podrían establecer la invariancia conductual

Una cuarta objeción sugiere que el argumento prueba demasiado. Si establece el principio de invariancia organizacional, un argumento similar establecería un principio de invariancia *conductual*. Para hacer esto, deberíamos construir un continuo de casos desde mi persona hasta cualquier sistema conductualmente equivalente. Se deduciría mediante un razonamiento similar que un sistema de esta clase debe ser consciente. Pero es plausible que algunos sistemas, como la tabla de doble entrada gigante de Block (1981) que almacena salidas para cada patrón de entradas, no sean conscientes. Por lo tanto, debe haber una falla en el argumento.

Esta objeción fracasa de dos modos. Primero, mi argumento se basó, en parte, en el hecho de que un sistema funcionalmente isomórfico tendrá la misma estructura cognitiva que yo y, en particular, los mismos juicios. Esto es lo que nos llevó a la conclusión de que el sistema desvaneciente José debía estar masivamente equivocado en sus juicios. El punto correspondiente no vale para sistemas conductualmente equivalentes. Un actor perfecto no tiene por qué poseer los mismos juicios que yo. Ni tampoco la tabla de doble entrada; ni los sistemas intermedios. Estos funcionarán mediante mecanismos relativamente diferentes.

Segundo, no es obvio en absoluto cómo podríamos ir desde mí a un isomorfo conductual arbitrario mediante pequeños pasos preservando la equivalencia conductual en todo momento. ¿Cómo haríamos esto para la tabla de doble entrada, por ejemplo? Quizás haya modos de hacerlo en grandes pasos cada vez, pero esto no será suficiente para el argumento: si hay pasos grandes entre sistemas vecinos, entonces ya no serían inverosímiles los qualia repentinamente desvanecientes. Con los isomorfos funcionales, había un modo natural de dar pasos muy pequeños, pero no hay un método natural semejante para los isomorfos conductuales. Por lo tanto, parece improbable que un argumento de este tipo pueda despegar.

A la postre, creo que el único modo defendible mediante el cual un oponente de la invariancia organizativa podría enfrentar este argumento es aceptar la posibilidad de los qualia gradualmente desvanecientes con la consiguiente posibilidad de que un sistema consciente racional pueda equivocarse masivamente acerca de su experiencia, o quizás aceptar los qualia repentinamente desvanecientes y las discontinuidades primitivas asociadas. Esta posición no es atractiva por su implicación de una disociación entre la conciencia y

la cognición, y la alternativa parece mucho más plausible; pero, a diferencia de otras objeciones, no es *obviamente* errónea. El argumento de qualia danzantes en el apartado 5 proporcionará aun más evidencia en contra de la posibilidad de los qualia ausentes, sin embargo, de modo que los oponentes del principio de invariancia no podrán descansar fácilmente.

Hago notar que un argumento parecido podría establecer que sistemas cuya organización funcional es *similar* (en lugar de idéntica) a la de un sistema consciente tendrán experiencias conscientes. El principio de invariancia tomado aisladamente es compatible con la tesis solipsista de que mi organización y sólo mi organización da origen a la experiencia. Pero podemos imaginar un cambio gradual en mi organización, así como nos imaginamos un cambio gradual en mi conformación física, bajo el cual mis creencias acerca de mi experiencia se preservarían fundamentalmente en todo momento; yo seguiría siendo un sistema racional, etc. Por razones similares a las ya enunciadas, parecería muy probable que la experiencia consciente se preserve en una transición de esta clase.

## 4. Qualia invertidos

El argumento de los qualia desvanecientes sugiere que mis isomorfos funcionales tendrán experiencias conscientes, pero no establece que los isomorfos tendrán el *mismo* tipo de experiencias conscientes. Es decir, la organización funcional determina la existencia o ausencia de la experiencia consciente, pero podría no determinar la naturaleza de esa experiencia. Para establecer que la organización funcional determina la naturaleza de la experiencia, deberemos establecer que los isomorfos funcionales con qualia *invertidos* son imposibles.

La idea de los qualia invertidos es familiar para la mayoría de nosotros. Pocas personas no se han preguntado en algún momento si lo que parece ser rojo para una persona no puede parecer azul para otra, y viceversa. Es uno de esos problemas filosóficos en los que al principio no estamos seguros de si la idea tiene sentido, y eso puede ser desconcertante incluso cuando se reflexiona sobre ello.

Aparentemente, la posibilidad de los qualia invertidos fue formulada por primera vez por John Locke en su *Ensayo acerca del entendimiento humano*:

> *Empero la idea de azul de una persona debería ser diferente de la de otra.* Tampoco significaría una imputación de falsedad a nuestras ideas simples, si debido a la diferente estructura de nuestros

> órganos ocurriese que *el mismo objeto debería producir en la mente de varios hombres diferentes ideas* al mismo tiempo; v.g. si la idea que una violeta produjo en la mente de un hombre por sus ojos fuese la misma que una caléndula produjo en la de otro hombre, y *viceversa.* Porque, como esto nunca podría saberse, debido a que la mente de un hombre no puede pasar al cuerpo de otro para percibir qué apariencias fueron producidas por esos órganos, y que ni las ideas involucradas, ni los nombres fueran confundidos en absoluto y ninguna falsedad estuviese en ninguno. Porque todas las cosas que tuviesen la textura de una violeta producirían constantemente la idea que llamó azul, y aquellas que tuviesen la textura de una caléndula producirían constantemente la idea que denominó amarillo, cualesquiera fueran esas apariencias en su mente, él podría ser capaz de distinguir regularmente cosas para su uso según esas apariencias y comprender y significar esas distinciones marcadas por los nombres "azul" y "amarillo", como si las apariencias o idea en su mente recibidas de esas dos flores fuesen exactamente las mismas que las ideas en la mente de los otros hombres. (libro 2, cap. 32, apart. 15)

Aquí Locke se ocupa de los qualia invertidos entre sistemas con una conducta similar, en lugar de entre isomorfos funcionales precisos. Parece expresar una posibilidad conceptual. La cuestión para nosotros es si expresa una posibilidad *empírica.*

Aun aquellos que se consideran a sí mismos materialistas supusieron con frecuencia que los isomorfos funcionales podrían tener experiencias conscientes diferentes. Por ejemplo, suele creerse que es naturalmente posible que un isomorfo funcional de mi persona con una conformación física diferente pueda tener experiencias de azul cuando yo tengo experiencias de rojo, o algo similar. Esta es la hipótesis de los qualia invertidos. Si es verdadera, entonces, aunque la existencia de la experiencia consciente podría sólo depender de la organización funcional, la naturaleza de las experiencias dependería de la conformación fisiológica o de algún otro factor no funcional.[10]

Vimos antes que es difícil mantener esta posición consistente con el materialismo. Si es naturalmente posible que un isomorfo funcional pueda tener qualia invertidos, entonces es lógicamente posible. Por lo tanto, también es lógicamente posible que un isomorfo *físico* tenga qualia invertidos, ya que no existe una conexión *conceptual* con un tipo específico de qualia ni para las neuronas ni para el silicio. Se deduce que la naturaleza de las experiencias específicas es un hecho ulterior que va más allá de los hechos físicos, y que el materialismo debe ser falso (a menos que aceptemos el enfoque de la "necesidad metafísica fuerte"). En lo que sigue, sin

embargo, dejaré este punto de lado. El análisis será independiente de la verdad del materialismo o del dualismo.

La posibilidad de los qualia invertidos, o del "espectro invertido" como a veces se lo conoce, se objeta algunas veces sobre la base verificacionista de que si ocurriese algo diferente nunca podríamos saberlo, de modo que no habría ninguna real diferencia (por ejemplo, Schlick, 1932). Obviamente no acepto estos argumentos: el mero hecho de que no podamos decir qué qualia experimenta un sistema no es suficiente para concluir que no exista una cuestión empírica, ya que la naturaleza de los qualia no está conceptualmente vinculada a la conducta; por lo tanto, dejaré de lado aquí esta objeción. Como analicé en el capítulo 3, a veces también se objeta la hipótesis sobre la base de que nuestro espacio de color es asimétrico, de modo que ninguna inversión podría definir una correspondencia apropiada (por ejemplo, Hardin, 1987; Harrison, 1967, 1973). Algunas de las respuestas que formulé antes siguen siendo apropiadas aquí, aun cuando la cuestión es ahora una de posibilidad natural; en particular, todavía podemos apelar a la posibilidad de una criatura con un espacio de color asimétrico y preguntarnos si podría tener un isomorfo funcional invertido. De cualquier forma, ignoraré esta preocupación y aceptaré, a los fines de la argumentación, que tenemos un espacio de color simétrico; sostendré que los qualia invertidos son de todas maneras imposibles.

El análisis de los qualia invertidos puede resultar confuso. Cuando digo "experiencia de azul", ¿quiero decir 1) lo que un *sujeto* llama una experiencia de "azul", 2) una experiencia causada por un objeto azul, o 3) lo que *yo* llamo una experiencia de "azul"? Elegiré el último uso. A lo largo de la exposición, por "experiencia de azul" querré decir el tipo de experiencia que *yo* llamo "azul", que tengo usualmente cuando veo cosas azules como el cielo o el mar, etc. En este uso, es concebible que otros (o incluso una versión futura de mí) puedan tener experiencias de azul causadas por objetos amarillos o por objetos que ellos llaman "rojos", etcétera.

Los argumentos en favor de los qualia invertidos suelen consistir simplemente en una aseveración de conceptibilidad, como con los qualia ausentes, pero estas aseveraciones claramente dejan abierta la cuestión de la posibilidad natural. Un par de argumentos en favor de la posibilidad natural de versiones de los qualia invertidos han sido formulados, pero ninguno amenaza el principio de invariancia organizacional.

El primero de estos argumentos defiende la posibilidad de qualia que estén invertidos mientras que la *conducta* se mantiene constante

(véase Gert, 1965; Lycan, 1973; y Wittgenstein, 1968). Primero, advertimos que los qualia podrían ser invertidos *dentro* de un sujeto, quizá mediante una reconexión de las conexiones de la retina a las áreas centrales en mi cerebro mientras estoy dormido. Cuando me despierto, afirmaré que el cielo repentinamente parece rojo, el pasto parece amarillo, etc., y tendré todo tipo de razones para creer que mis qualia han sido invertidos. Luego, planteamos la posibilidad de alguien cuyo cerebro fue reconectado de este modo desde el nacimiento. Una persona así podría tener qualia que estén sistemáticamente invertidos respecto de la norma pero, por supuesto, habrá aprendido a llamar al cielo azul, al pasto verde, etc., de manera que esta inversión podría no surgir nunca en su conducta.[11]

Sin embargo, esto no establece la posibilidad de los qualia invertidos con una organización funcional fija. Para ver esto sólo necesitamos notar que al reconectar mis conexiones, el demonio *cambió* mi organización funcional de un modo significativo. De la misma manera, la organización funcional del sujeto que fue reconectado desde el nacimiento se alteró, de modo que él no será un duplicado funcional mío. Podría compartir algunas de mis propiedades funcionales en un nivel poco detallado, pero seguramente no compartirá mi organización funcional detallada. El principio de invariancia organizacional, por lo tanto, no está amenazado por estos casos.[12]

Un argumento relacionado, formulado por Block (1990), concierne a la "Tierra Invertida", en la que el cielo es amarillo, el pasto es rojo, etc.[13] Debemos suponer que soy secuestrado y llevado a la Tierra Invertida, pero al mismo tiempo se me dan lentes de contacto que invierten los colores, de modo que todo me parece normal. Block utiliza este escenario para oponerse a un enfoque representacionalista de los qualia donde, por ejemplo, una experiencia de azul es equiparada a un estado perceptual acerca de cosas azules (después de algún tiempo en Tierra Invertida, nuestras experiencias de azul serán acerca de cosas *amarillas*). También lo utiliza para oponerse a un enfoque funcionalista del tipo en el cual las experiencias de azul se equiparan a los estados causados por objetos azules.

Nuevamente, este caso no influye en la refutación del principio de invariancia organizacional. Después de todo, cuando veo el cielo amarillo a través de mis lentes invertidos en Tierra Invertida, mi organización funcional interna será exactamente la misma que cuando veo el cielo azul en la Tierra y mi experiencia también será la misma, tal como el principio lo predice. En el mejor de los casos, el experimento mental disocia las experiencias de las propiedades de nuestro *ambiente* y de las propiedades funcionales "amplias" que involucran a nuestro ambiente; pero la organización funcional de la

que nos ocupamos es totalmente interna. El experimento mental no produce ningún caso en el que dos sistemas internamente isomórficos tengan experiencias diferentes, de modo que el principio de invariancia no resulta afectado.[14]

## 5. Qualia danzantes

Sería posible pensar que podríamos directamente adaptar el argumento de los qualia desvanecientes para producir un argumento en contra de la posibilidad de los qualia invertidos. Desafortunadamente esto no funcionará. Imagínese cómo sería un argumento análogo. Partimos de mi persona que está teniendo una experiencia de rojo y de un sistema invertido que tiene una experiencia de azul. Mediante un reemplazo gradual, construimos una serie de casos, cada uno de los cuales experimenta algún color intermedio. ¡Pero no hay nada malo en esto! Los sistemas intermedios son simplemente casos de inversión leves de qualia y no son más problemáticos que el caso extremo.

Podría no ser obvio qué es lo que exactamente experimentan los sistemas intermedios. Tal vez ningún color de nuestro espacio usual de colores pueda hacer el trabajo de un modo consistente con los patrones de categorización y diferencias del sistema. Quizás experimenten colores enteramente nuevos, colores que yo no puedo experimentar pero que, no obstante, forman un continuo desde el rojo al azul. Esto sería muy extraño, pero no es del todo inverosímil. Lo importante es que el problema que encontramos en el caso de los qualia desvanecientes está aquí totalmente ausente. Estos sistemas *no* estarán sistemáticamente equivocados acerca de las características de su experiencia. Cuando afirman experimentar distinciones, podrían estar experimentándolas; cuando afirman experiencias intensas, tienen experiencias intensas, etc. Seguramente, los colores que llaman "rojo" serán diferentes de los que yo llamo rojo, pero esto no es problemático; ya ocurre en el caso usual de inversión. Lo importante es que, a diferencia del caso de los qualia desvanecientes, las características *estructurales* de las experiencias de estos sistemas se preservan en todo momento.

Sin embargo, es posible encontrar en las cercanías un buen argumento en contra de la posibilidad de los qualia invertidos.[15] Una vez más, para los propósitos de la *reductio*, supóngase que los qualia invertidos son empíricamente posibles. Puede haber dos sistemas funcionalmente isomórficos en el mismo estado funcional pero con experiencias diferentes. Supóngase, con fines ilustrativos, que estos sistemas somos yo mismo, teniendo una experiencia de rojo, y mi

isomorfo de silicio, teniendo una experiencia de azul (existe una pequeña salvedad acerca de la generalidad, que analizaré más adelante).

Como antes, construimos una serie de casos intermedios entre mi persona y mi isomorfo. Aquí, el argumento toma un giro diferente. No necesitamos preocuparnos acerca del *modo* en el cual las experiencias cambian cuando nos movemos a lo largo de la serie. Tal vez cambien repentinamente, quizá salten a través de todo el mapa, aunque seguramente es más plausible que cambien en forma gradual. Todo lo que importa es que debe haber dos puntos A y B en esta serie, tal que 1) no se reemplace más de un décimo del cerebro entre A y B, y 2) A y B tienen experiencias significativamente diferentes. Para ver que este debe ser el caso, sólo necesitamos considerar los puntos en los cuales el 10%, el 20%, etc., hasta el 90% del cerebro ha sido reemplazado. Las experiencias de rojo y azul son suficientemente diferentes de modo que algunos pares vecinos *deben* ser significativamente diferentes (esto es, lo suficientemente diferentes para que la diferencia sea notable si fueran experimentados por la misma persona); no hay ninguna manera de ir desde el rojo al azul mediante diez saltos imperceptibles.

Es verdad que puede haber diferencias imperceptibles entre experiencias diferentes. Si cambiamos un tono de rojo suficientemente poco, no podremos percibir la diferencia. Podríamos suponer que esto se debe a que no hay una diferencia en la experiencia, sólo una diferencia en el mundo; pero si esto fuese todo lo que ocurre podríamos iterar un cambio de este tipo mil veces y eventualmente mostraríamos que el rojo y el azul producen las mismas experiencias, lo que es ridículo. De modo que puede haber *alguna* diferencia en la experiencia que no sea perceptible. Podemos observar este fenómeno si observamos una amplia extensión de pintura de tonos sutilmente variables; a veces es sumamente difícil decir si nuestras experiencias de diferentes partes es la misma o diferente. Pero, lo importante es que las diferencias imperceptibles son muy *pequeñas*. Cuanto más, diez de estos saltos nos podrían llevar desde un tono de rojo a un tono sutilmente diferente del mismo color. (Esto abre una pequeña falla en la generalidad del argumento; volveré a este punto más adelante.)

Entre los sistemas de rojo y azul debe haber, por lo tanto, dos sistemas que difieren cuanto más en un 10% de su conformación interna, pero que tienen experiencias significativamente diferentes. Con fines ilustrativos, hagamos que estos sistemas seamos Guillermo y yo mismo. Cuando yo tengo una experiencia de rojo, Guillermo tiene una experiencia levemente diferente. Podríamos también suponer que Guillermo ve azul; quizá su experiencia sea más similar a la mía

que eso, pero no hace ninguna diferencia al argumento. Los dos sistemas también difieren en que allí donde hay neuronas en alguna pequeña región de mi cerebro hay chips de silicio en el cerebro de Guillermo. Esta sustitución de un circuito de silicio por un circuito neuronal es la única diferencia física entre Guillermo y mi persona.

El paso crucial en el experimento mental es tomar un circuito de silicio como el de Guillermo e instalarlo en mi propia cabeza como un *circuito de respaldo*. Este circuito será funcionalmente isomórfico a un circuito que ya existe en mi cabeza. Equipamos el circuito con transductores y efectores para que pueda interactuar con el resto de mi cerebro, pero no lo conectamos directamente. En cambio, instalamos un *conmutador* que pueda alternar directamente entre los circuitos neuronales y los de silicio. Al accionar hacia un lado el conmutador, el circuito neuronal se vuelve irrelevante y el circuito de silicio asume el control. Podemos imaginar que el conmutador controla los puntos de interfaz en los que los circuitos pertinentes afectan el resto del cerebro. Cuando el conmutador está hacia un lado, las conexiones del circuito neuronal se dejan de lado y se conectan los efectores del circuito de silicio. (Podemos imaginarnos que los transductores para los dos circuitos están conectados todo el tiempo, de modo que el estado de ambos circuitos evoluciona apropiadamente, pero sólo un circuito por vez está involucrado en el procesamiento. Podríamos también realizar un experimento similar en el que los transductores y efectores están desconectados, para asegurarnos de que el circuito de respaldo esté totalmente aislado del resto del sistema. Esto cambiaría unos pocos detalles, pero la moraleja seguiría siendo la misma.)

Inmediatamente después de accionar el conmutador, el procesamiento que un momento antes era realizado por el circuito neuronal es ahora realizado por el circuito de silicio. El flujo de control dentro del sistema fue redirigido. Sin embargo, mi organización funcional es exactamente la misma que la que habría sido si no hubiésemos accionado el conmutador. La única diferencia significativa entre los dos casos es la conformación física de un circuito dentro del sistema. También existe una diferencia en la conformación física de otro circuito "ocioso", pero esto es irrelevante para la organización funcional, ya que no tiene ningún papel en afectar otros componentes del sistema y dirigir la conducta.

¿Qué ocurre con mi experiencia cuando acciono el conmutador? Antes de instalar el circuito, yo experimentaba el rojo. Después de instalado pero antes de accionar el conmutador, podemos suponer que todavía seguiré experimentando el rojo, ya que la única diferencia es

la adición de un circuito que no está de ninguna manera involucrado en el procesamiento; por la significación que tiene para mi procesamiento, bien me lo podría haber comido. *Después* de accionar el conmutador, sin embargo, soy más o menos el mismo sistema que Guillermo. La única diferencia entre Guillermo y yo ahora es que yo tengo un circuito neuronal causalmente irrelevante colgando del sistema (podríamos imaginarnos que el circuito se destruye cuando se acciona el conmutador). Guillermo, por hipótesis, disfrutaba una experiencia de azul. Después de la conmutación, entonces, yo también experimentaré el azul.

Lo que ocurrirá, entonces, es que mi experiencia cambiará "ante mis ojos". Donde yo antes experimentaba el rojo, ahora experimento el azul. De pronto, tendré una experiencia de *azul* de la manzana que está sobre mi escritorio. Podemos incluso imaginarnos accionando el conmutador hacia un lado y hacia el otro un número de veces, de modo que las experiencias de rojo y azul "dancen" ante mis ojos.

Esto podría parecer razonable al principio —es una imagen extrañamente atractiva—, pero ocurre algo muy extraño. Mis experiencias pasan del rojo al azul, pero *yo no noto ningún cambio*. Aun mientras accionamos el conmutador un número de veces y mis qualia danzan hacia un lado y hacia el otro, yo simplemente sigo haciendo mis cosas sin notar nada inusual. Por hipótesis, mi organización funcional sigue siendo normal en todo momento. En particular, mi organización funcional después de accionar el conmutador evoluciona exactamente como lo habría hecho si no se lo hubiese accionado. No existe ninguna diferencia especial en mis disposiciones conductuales. No estoy repentinamente dispuesto a decir "¡Hmmm! Algo extraño está sucediendo!" No hay espacio para un sobresalto repentino, para una exclamación, o incluso para una distracción de la atención. Cualquier reacción inusual implicaría una diferencia funcional entre los dos circuitos, lo que sería contrario al isomorfismo estipulado. Por una cuestión de diseño, mi organización cognitiva es exactamente como solía ser y, en particular, es precisamente como hubiese sido si el conmutador no hubiese sido accionado.

Ciertamente, bajo cualquier interpretación funcional de las creencias, es evidente que no puedo adquirir ninguna nueva creencia mientras ocurre la conmutación. Aunque cuestionemos una concepción funcional, es extremadamente inverosímil que un simple reemplazo de un circuito neuronal por un circuito de silicio que preserva la organización general pueda ser responsable de la adición de nuevas creencias significativas como "Mis qualia acaban de cambiar". Como en el caso de los qualia desvanecientes, simplemente no hay espacio

para que ocurra un cambio de este tipo, a menos que sea en una mente incorpórea cartesiana acompañante.

Por lo tanto nos vemos llevados una vez más a una *reductio ad absurdum*. Parece totalmente inverosímil suponer que mis experiencias puedan cambiar de un modo tan significativo, mientras les presto una cuidadosa atención, sin que yo pueda advertir el cambio. Esto sugeriría, nuevamente, una disociación radical ente la conciencia y la cognición. Si este tipo de cosas pudiera ocurrir, entonces la psicología y la fenomenología estarían radicalmente fuera de sincronía; mucho más fuera de sincronía de lo que incluso el escenario de los qualia desvanecientes podría sugerir.

Este escenario de "qualia danzantes" podría ser lógicamente posible (aunque el caso es tan extremo que parece *apenas* lógicamente posible), pero eso no significa que sea plausible como posibilidad empírica; no más plausible que el mundo haya sido creado hace cinco minutos. Como hipótesis empírica, parece mucho más plausible que, cuando nuestras experiencias cambian de manera significativa, entonces, en tanto seamos racionales y prestemos atención, deberíamos poder advertir el cambio. De no ser así, la conciencia y la cognición estarían conectadas tan sólo por el más delgado de los hilos.

Si suponemos que los qualia danzantes son naturalmente posibles, nos vemos llevados a un pensamiento preocupante: podrían ser *reales* y estar ocurriéndonos todo el tiempo. Las propiedades fisiológicas de nuestros mecanismos funcionales cambian constantemente. Las propiedades funcionales de los mecanismos son razonablemente robustas; esperaríamos que esta robustez esté asegurada por la evolución. Pero no hay ninguna razón adaptativa para que las propiedades no funcionales se mantengan constantes. De un momento a otro ciertamente existirán cambios en las propiedades moleculares de bajo nivel. Propiedades como la posición, la conformación atómica, etc., pueden cambiar mientras que el papel funcional se preserva, y esos cambios casi seguramente ocurren de modo permanente.

Si permitimos que los qualia dependan no sólo de la organización funcional sino también de los detalles de implementación, bien podría ocurrir que *nuestros* qualia estén de hecho danzando ante nuestros ojos todo el tiempo. No parece haber ninguna razón de principio para que un cambio de neuronas por silicio deba marcar una diferencia, mientras que un cambio en la realización neuronal no deba hacerlo;[16] el único lugar en el cual trazar una línea *basada en principios* es el nivel funcional.[17] La razón de que dudemos de que esa danza ocurra en nuestro propio caso es que aceptamos el siguiente principio: cuando nuestra experiencia cambia significativamente, podemos

notar el cambio. Si aceptásemos la posibilidad de los qualia danzantes en el caso original, estaríamos desechando este principio, y ya no estaría disponible como defensa en contra del escepticismo aun en los casos más usuales.

No está fuera de cuestión que podamos verdaderamente realizar este experimento. Desde ya, las dificultades prácticas serían inmensas pero, al menos en principio, podríamos instalar un circuito de este tipo en mí y *yo* podría ver qué ocurre e informarlo al mundo. No obstante, no tiene sentido realizar el experimento; sabemos cuál será el resultado. Yo informaré que mi experiencia se mantuvo igual en todo momento, un tono constante de rojo, y que no advertí nada extraño. Yo estaré aun más convencido que antes de que los qualia están determinados por la organización funcional. Por supuesto, esto no será una *prueba*, pero será difícil cuestionar la evidencia seriamente.

Concluyo que, de lejos, la hipótesis más plausible es que un reemplazo de neuronas que preserve la organización funcional preservará los qualia. Los problemas con el escenario de los qualia danzantes pueden atribuirse al supuesto inicial de que un sistema de silicio funcionalmente isomórfico podría experimentar el azul donde yo experimento el rojo. La reacción más razonable es eliminar este supuesto y concluir que la experiencia está totalmente determinada por la organización funcional.

Debe observarse que este experimento mental funciona tan bien en contra de la posibilidad de los qualia ausentes como en contra de los qualia invertidos. Simplemente tomamos dos puntos en el camino hacia los qualia ausentes entre los cuales la experiencia difiere significativamente e instalamos un circuito de respaldo de la misma manera. Como antes, si los qualia ausentes son posibles, entonces la conmutación causará que mis qualia oscilen ante mis ojos, de vivo a pálido y de vuelta, sin que yo nunca lo advierta. Nuevamente, es mucho más plausible que esta danza imperceptible sea imposible, de modo que los qualia ausentes son imposibles.

Personalmente, encuentro que este es un argumento aun más convincente en contra de los qualia ausentes que el argumento del apartado 3, aunque ambos tienen un papel que desempeñar. Un opositor podría sencillamente aceptar la posibilidad de los qualia desvanecientes, pero la dificultad de aceptación de los qualia danzantes parece ser de un orden de magnitud mayor. La propia inmediatez de la conmutación parece hacer una diferencia significativa, como también el hecho de que el fenómeno que el sujeto no puede percibir es tan dinámico y sorprendente. Los qualia desvanecientes significarían

que algunos sistemas están fuera de contacto con su experiencia consciente, pero los qualia danzantes establecerían una brecha aún más extraña.

Debido a la estructura del argumento de los qualia danzantes, este está expuesto a algunas pocas vulnerabilidades más que el argumento de los qualia desvanecientes. Sin embargo, no parece que ninguna de ellas pueda ser utilizada para llevar a un opositor a una posición ventajosa. En lo que sigue analizaré estas vulnerabilidades, junto con otra objeción al argumento. Pueden volverse a formular nuevamente todas las objeciones que mencionamos para el caso de los qualia desvanecientes y las respuestas serán más o menos las mismas, de modo que no me molestaré en repetirlas aquí.

### Objeción 1: Vulnerabilidades acerca de la velocidad y la historia

El argumento que formulé aquí puede extenderse naturalmente del caso neuronal/silicio a muchos otros ejemplos de isomorfos funcionales, pero hay un par de excepciones que involucran la velocidad y la historia. Si un isomorfo es mucho más veloz o más lento que el sistema original, no podemos simplemente sustituir un circuito de un sistema por el otro y esperar que todo funcione normalmente. De modo que el argumento, tal cual lo formulé, no descarta la posibilidad de que un cambio en la velocidad que mantenga constante la organización funcional pueda ser responsable de una inversión en los qualia. Una deficiencia similar queda abierta para los isomorfos físicos que difieren meramente en su *historia*: tal vez, si yo hubiese nacido en el hemisferio sur, experimentaría el verde, mientras que un gemelo físico nacido en el hemisferio norte experimentaría el rojo. La historia no puede variarse en un escenario de qualia danzantes (aunque puede variarse en un escenario de qualia desvanecientes), de modo que el argumento no tiene relación con la hipótesis de que los qualia supervienen al pasado.

Pero ninguna de estas hipótesis era muy plausible en primer lugar. Es razonable que la historia afecte nuestros qualia al afectar nuestra estructura física, pero la dependencia de la historia que requerimos más arriba sería mucho más fuerte: habría un efecto "no local" de la historia distal sobre los qualia actuales, no mediado por ninguna cosa en la estructura física o proximidad en el espacio y el tiempo. En lo que respecta a la velocidad, parecería bastante arbitrario que un cambio en ella pueda invertir los qualia cuando ninguna otra cosa puede hacerlo. Estas hipótesis son coherentes, de modo que un opositor *podría* aceptarlas, pero hay pocas razones para hacerlo.

Una vez que hemos establecido que todos los otros cambios que preservan la organización preservan los qualia, hay poco atractivo en la idea de que la velocidad o la historia podrían ser las únicas cosas que hacen una diferencia.

### Objeción 2: ¿Qué hay de las inversiones leves?

Otra pequeña vulnerabilidad es que el argumento no refuta la posibilidad de inversiones muy leves del espectro. Entre el rojo oscuro y un rojo levemente más oscuro, por ejemplo, podría haber nueve tonos intermedios tal que ningún par de tonos vecinos sea distinguible. En este caso el escenario de los qualia danzantes no es un problema; si el sistema no advierte ninguna diferencia al accionar el conmutador, eso es justamente lo que esperaríamos.

Por supuesto, no hay nada especial en la cifra de un décimo como magnitud de la diferencia entre dos sistemas vecinos. Pero no podemos aumentar demasiado la cifra. Si la aumentásemos hasta un medio, encontraríamos problemas con la identidad personal: podría ser razonable sugerir que al accionar el conmutador creamos una nueva persona, y no sería un problema que la nueva persona no advierta ningún cambio. Tal vez pudiésemos llegar hasta el 20 o 25% sin esos problemas; pero aun esto permitiría la posibilidad de inversiones muy leves, del tipo que podría componerse de cuatro o cinco cambios imperceptibles. Podemos reducir el impacto de esta preocupación, sin embargo, si notamos que es muy improbable que la experiencia dependa por igual de todas las áreas del cerebro. Si la experiencia del color depende fundamentalmente de una pequeña área de la corteza visual, digamos, entonces podríamos realizar cualquier inversión de qualia de un solo golpe mientras sólo reemplazamos una pequeña porción del sistema, y el argumento tendría éxito incluso en contra de la inversión del qualia más leve perceptible.

De todas maneras, cualquier vulnerabilidad de este tipo no es peligrosa. En el peor de los casos, dejamos abierta la posibilidad de una subdeterminación extremadamente leve de la experiencia por parte de la organización. Este tipo de subdeterminación podría parecer tan leve que no resultaría interesante pero, de todas formas, podemos advertir que lleva a una posición poco atractiva. Parecería razonable que las experiencias deban ser invertibles en todos los casos o no invertibles en absoluto, sin embargo, ¿por qué el mundo debería ser tal que es posible una pequeña inversión pero nada más? Esto parecería bastante arbitrario. No podemos descartarlo, pero no es una hipótesis muy plausible.

## Objeción 3: Qualia no atendidos

De un modo similar, el argumento deja abierta la vulnerabilidad de que los qualia *no atendidos* podrían ser invertibles. Si no atendemos a la periferia de nuestro campo visual, por ejemplo, una inversión de qualia podría ocurrir allí sin que lo notemos. Experimentos recientes (Rensink, O'Regan y Clark, 1995) muestran que podemos modificar características bastante significativas de una imagen que un sujeto está mirando sin que este lo advierta, si no está concentrándose justamente en esas características (estos experimentos por lo general involucran un intervalo temporal breve entre la exhibición de dos imágenes, de manera que no es del todo similar al escenario de los qualia danzantes, pero se acerca). Estos argumentos, entonces, dejan abierta la posibilidad de que los qualia *no atendidos* puedan ser invertibles.

Sin embargo, nada en este tipo de consideración sugiere que los qualia *no atendidos* puedan ser invertibles. De modo que aprovechar esta vulnerabilidad nos dejaría en la posición poco atractiva de que los qualia son organizacionalmente invariantes cuando son lo bastante centrales en nuestra atención, pero dependientes de otras características cuando no lo son. (¿Es posible que una experiencia de verde invertida en la periferia pueda volver al rojo cuando no la atendemos?) Parece muy improbable que una posición de este tipo pueda hacerse teóricamente satisfactoria. Como con las otras vulnerabilidades, esta sólo abre el camino a posiciones que carecen de cualquier plausibilidad significativa.

## Objeción 4: Doble conmutación

Otra objeción es la siguiente.[18] Podemos imaginarnos un experimento relacionado en el cual reconectamos las conexiones de las entradas de rojo y azul a las áreas centrales del cerebro de modo que las entradas de azul desempeñen el papel que las entradas del rojo desempeñaban antes, y en las cuales también reconectamos sistemáticamente las conexiones *descendentes* desde el área central para compensar. Cuando una entrada de azul causa que el área central ingrese en un estado previamente asociado con el rojo, las conexiones desde el área central al resto del cerebro son reconectadas de modo que el resto del cerebro funcione tal como lo habría hecho si no hubiese habido ninguna reconexión. De este modo, mi experiencia casi seguramente conmutará del rojo al azul, pero mis disposiciones conductuales permanecerán constantes en todo momento. En este caso, una conmutación repetida seguramente llevaría a qualia danzantes. ¿Los qualia danzantes no serían razonables después de todo?

Primero, debe observarse que esta reconexión sería una tarea mucho más vasta que en cualquiera de los otros casos que describí. El área central afectará el resto del cerebro en todo tipo de lugares diferentes. Cada una de estas conexiones deberá ser reconectada y, crucialmente, ninguna reconexión simple podrá hacer la tarea en ninguno de dichos lugares. No podemos simplemente conmutar “salidas de rojo” en “salidas de azul”, como podríamos hacer con las entradas; las salidas desde el sistema central pueden representar cosas tan diversas como recuerdos recuperados, instrucciones motoras, etc., las que podrían no tener ninguna diferencia simple de “polaridad” entre una salida de rojo y una salida de azul. Para determinar una “salida de azul” apropiada, probablemente necesitaríamos simular todo el procesamiento del área central, dado su estado inicial y entrada, para ver qué produce. Si esto es así, será la simulación la que haga el trabajo causal, no la propia área central, y la fuerza del escenario se perderá.

Segundo, aun si existiese un modo simple en el que las salidas pudieran ser reconectadas, nótese que *sólo* las disposiciones conductuales se preservan, no la organización funcional. ¿Cómo podría experimentarse esto? En este caso, imagino que yo *advertiría* la conmutación e intentaría actuar en forma acorde, pero me sentiría como si algún irritante titiritero estuviese interfiriendo con mis acciones. A diferencia del caso previo, habrá *espacio* para estas creencias extra y otros estados cognitivos; estos estarán sustentados por los diferentes estados del área central. Podemos imaginar que una vez que ocurre la retroalimentación y la entrada a las áreas centrales indica que sus movimientos motores fueron totalmente diferentes de lo que se planeó, podemos imaginar que el estado del área central será severamente alterado. De hecho, esto nos lleva de vuelta a la primera objeción, ya que parecería casi imposible poder compensar sistemáticamente esos efectos de la retroalimentación. En cualquier caso, la diferencia significativa en la organización funcional revela que los casos no son análogos.

## 6. Funcionalismo no reductivo

En síntesis: hemos establecido que si los qualia ausentes son posibles, entonces los qualia desvanecientes también lo son; si los qualia invertidos son posibles, entonces los qualia danzantes son posibles; y si los qualia ausentes son posibles, entonces los qualia danzantes también lo son. Pero es poco razonable que los qualia desvanecientes sean posibles y es extremadamente inverosímil que los qualia

danzantes lo sean. Por lo tanto, es extremadamente inverosímil que los qualia ausentes y los qualia invertidos sean posibles. Se deduce que tenemos buenas razones para creer que el principio de invariancia organizacional es verdadero y que la organización funcional determina completamente la experiencia consciente.

Debe notarse que estos argumentos no establecen el funcionalismo en el sentido más fuerte ya que, cuanto más, establecen que los qualia ausentes e invertidos son empíricamente (o naturalmente) imposibles. Hay dos razones de que los argumentos no puedan extenderse en argumentos a favor de la imposibilidad *lógica* de los qualia ausentes e invertidos, como a algunos funcionalistas podría gustarle. Primero, los qualia desvanecientes y los qualia danzantes parecen ser hipótesis coherentes, pero no son plausibles. Algunos podrían cuestionar la posibilidad lógica de estas hipótesis; tal vez, se pueda sostener que es constitutivo de los qualia que podamos advertir las diferencias en ellos. Esta intuición conceptual es cuestionable pero, en cualquier caso, existe una segunda razón de que estos argumentos no logren establecer la determinación lógica de la experiencia por parte de la organización funcional.

Para ver esta segunda razón, nótese que los argumentos adoptan como premisa *empírica* ciertos hechos acerca de la distribución de la organización funcional en los sistemas físicos: que tengo experiencias conscientes de una cierta clase o que algunos sistemas biológicos las tienen. Si estableciésemos la imposibilidad lógica de los qualia desvanecientes y danzantes, esto podría establecer la necesidad lógica del *condicional*: si un sistema con organización funcional de grano fino $F$ tiene un cierto tipo de experiencias conscientes, entonces cualquier sistema con la organización $F$ tiene esas experiencias. Pero no podemos establecer la necesidad lógica del consecuente de ese condicional sin establecer la necesidad lógica de la premisa, y la premisa es ella misma empírica. Para establecer la determinación lógica de la experiencia por parte de la organización funcional, primero deberíamos establecer la superveniencia lógica de la experiencia a lo físico, algo que argumenté que no puede hacerse. Aunque *pudiésemos* establecer la superveniencia lógica a lo físico, sería probablemente mediante una definición funcional, pero con una definición de este tipo la imposibilidad lógica de los qualia ausentes e invertidos se deduciría sin ninguna necesidad de argumentos caprichosos. De cualquier forma, entonces, los argumentos de los qualia desvanecientes y danzantes son de poca utilidad para defender la imposibilidad lógica o metafísica de los qualia ausentes o invertidos.

Los argumentos, por lo tanto, fracasan en establecer una forma fuerte de funcionalismo en la cual la organización funcional es

*constitutiva* de la experiencia consciente; pero tienen éxito en establecer la forma más débil que denominé *funcionalismo no reductivo*, según la cual la organización funcional es suficiente para la experiencia consciente con necesidad natural. Según este enfoque, la experiencia consciente está determinada por la organización funcional, pero no tiene por qué ser reducible a la organización funcional.

De cualquier manera, la conclusión todavía es fuerte. El principio de invariancia nos dice que, en principio, los sistemas cognitivos que se realizan en cualquier tipo de medio pueden ser conscientes. En particular, la conclusión apoya firmemente las ambiciones de los investigadores en inteligencia artificial, como analizaré más detenidamente en el capítulo 9. Si el funcionalismo no reductivo es correcto, la irreducibilidad de la conciencia no levanta ninguna barrera a la eventual construcción de un mecanismo computacional consciente.

Lo que es muy importante, hemos avanzado en nuestro intento por restringir los principios en virtud de los cuales la conciencia naturalmente superviene a lo físico. Hemos restringido las propiedades significativas en la base de superveniencia a las propiedades *organizacionales*. En un cierto sentido, podemos decir que no sólo la conciencia superviene a lo físico, sino que también superviene a lo organizativo. Esto debe especificarse cuidadosamente, debido al hecho de que todo sistema realiza numerosos tipos de organización funcional, pero podemos decir lo siguiente: para todo sistema físico que da origen a la experiencia consciente, existe alguna organización funcional *F* realizada por el sistema, tal que es naturalmente necesario que cualquier sistema que realice *F* tenga experiencias conscientes idénticas. Para seleccionar la *F* relevante, debemos emplear un grano suficientemente fino que nos permita fijar los estados cognitivos como los juicios. Esto, a su vez, puede lograrse requiriendo que *F* sea de un grano suficientemente fino como para fijar los mecanismos responsables de la producción de la conducta y fijar las disposiciones conductuales. Esto es todo lo que los argumentos de los qualia desvanecientes y danzantes requerían, de modo que es todo lo que necesitamos para la invariancia organizacional.

Por lo tanto, es una ley para ciertas organizaciones funcionales *F* que la realización de *F* estará acompañada por un tipo específico de experiencia consciente. Esto no significa que sea una ley *fundamental*. Sería extraño que el universo tuviese leyes fundamentales que conectasen organizaciones funcionales complejas con experiencias conscientes. Más bien, esperaríamos que fuese una consecuencia de leyes psicofísicas más simples y fundamentales. Mientras tanto, el principio de invariancia organizacional podrá actuar como una fuerte restricción sobre una teoría última.

# 8

# Conciencia e información: algo de especulación

## 1. Hacia una teoría fundamental

Hasta ahora, hemos aislado unas pocas conexiones entre la conciencia y los procesos físicos que merecen ser llamadas leyes psicofísicas. Una de estas es el principio de coherencia que conecta la conciencia con la percatación, o la disponibilidad global. Otra es el principio más específico de la coherencia estructural, que conecta la estructura de la conciencia con la estructura de la percatación. El principio de invariancia organizacional es el tercero. Estos principios pueden ser *componentes* de una teoría final de la conciencia. Nos permiten utilizar hechos físicos para predecir e incluso explicar ciertos hechos acerca de la experiencia consciente. También, *constriñen* la forma de una teoría final de la conciencia: si una teoría de este tipo no es compatible con esas leyes, es improbable que sea correcta. Pero debe haber más que esto. Estos principios no alcanzan para constituir una teoría final ni nada que se le parezca.

El problema es que ninguno de ellos es un candidato plausible para una ley *fundamental* de una teoría de la conciencia. Todos ellos expresan regularidades en un nivel relativamente alto. El concepto de percatación (o disponibilidad global) es un concepto de alto nivel, por ejemplo, y sus límites son algo vagos; es muy improbable que este concepto pueda estar involucrado en una ley fundamental. El principio de invariancia organizacional podría ser menos vago, pero todavía expresa una regularidad en un nivel que está lejos del fundamental. Otro problema: estos principios subdeterminan ampliamente la naturaleza de la conexión psicofísica. Todo tipo de preguntas acerca de la conexión permanecen sin respuesta. Por ejemplo, ¿exactamente

qué *clase* de organización da origen a la experiencia consciente? ¿Cuán simple puede ser una organización antes de que la experiencia desaparezca? Y, ¿cómo podemos predecir el carácter específico de una experiencia (no sólo su estructura) a partir de su base física? Nos gustaría una teoría completa de la conciencia para responder estas preguntas, pero los principios estudiados hasta ahora no ayudan.

Para una teoría final, necesitamos un conjunto de leyes psicofísicas análogas a las leyes fundamentales de la física. Estas leyes fundamentales (o *básicas*) deberán formularse en un nivel que conecte las propiedades básicas de la experiencia con características simples del mundo físico. Las leyes deberían ser precisas y, en conjunto, no deberían dejar espacio alguno para la subdeterminación. Cuando se las combina con los hechos físicos acerca de un sistema, deberían permitirnos predecir perfectamente los hechos fenoménicos acerca del mismo. Además, así como las leyes básicas de la física implican todas las leyes y regularidades físicas de nivel superior (por lo menos cuando se las combina con condiciones fronterizas), las leyes básicas acerca de la conciencia deben implicar y explicar las diversas leyes no básicas, como los principios de coherencia y el principio de invariancia organizacional. Una vez que hemos formulado las leyes físicas *y* psicofísicas, podemos en cierto sentido comprender la estructura básica del universo.

Esta es una tarea difícil y no la alcanzaremos en un futuro cercano. Pero, al menos, podemos avanzar en esa dirección. Los principios de invariancia organizacional y coherencia estructural ya plantean una fuerte restricción sobre la forma de una teoría fundamental, y no existe un vasto número de candidatos para construcciones básicas que pudiesen ser los ingredientes fundamentales de la teoría. En este capítulo, presentaré algunas ideas para una teoría fundamental. No presento una teoría completamente desarrollada con un conjunto amplio de leyes básicas, pero formulo sugerencias acerca de las construcciones involucradas en esas leyes y de cuál podría ser su forma general. Podríamos considerarla una *prototeoría*: un armazón en torno del cual podría construirse una teoría.

Las ideas en este capítulo serán mucho más esquemáticas y más especulativas que las que presentamos en otros lugares del libro; plantean tantas respuestas como preguntas. Es muy probable, también, que estén totalmente equivocadas. Mi objetivo al formular estas ideas sueltas no es plantear un marco conceptual que resista un escrutinio filosófico detallado; más bien, las propongo con el fin de poner las ideas sobre la mesa. Tenemos que *comenzar* a pensar acerca de las teorías fundamentales de la conciencia, y quizás haya algo útil aquí que pueda desarrollarse.

## 2. Aspectos de la información

La noción básica de la que me ocuparé en este capítulo es la de *información.* Existen muchos diferentes conceptos de información a la deriva en el espacio de las ideas contemporáneas, de modo que lo primero que tenemos que hacer cuando hablamos acerca de la información es aclarar de qué estamos hablando. El concepto de información del que me ocuparé tiene mucho en común con el formulado por Shannon (1948). Aquí, presentaré una adaptación y desarrollo de esta idea. Mantendré el desarrollo en un nivel relativamente informal; sólo proporcionaré el nivel de formalismo necesario para capturar los aspectos más fundamentales del concepto que sean pertinentes. Existen algunos pocos tecnicismos en este apartado, pero los apartados siguientes son más sencillos.

Shannon no se ocupó de una noción semántica de la información según la cual esta es siempre información *acerca* de algo. Más bien, se concentró en una noción formal o sintáctica en la que la clave es el concepto de un estado seleccionado a partir de un conjunto de posibilidades. El tipo más básico de información es el *bit*, que representa una elección entre dos posibilidades: un solo bit (0 o 1) seleccionado de un espacio de dos estados se dice que contiene información. En un caso más complejo, un "mensaje" como "0110010101" seleccionado de un espacio de posibles mensajes binarios contiene información de un modo similar. Lo importante, en la concepción de Shannon, no es cualquier *interpretación* de esos estados; lo significativo es la *especificidad* de un estado dentro de un espacio de diferentes posibilidades.

Podemos formalizar esta idea mediante el concepto de *espacio de información.* Un espacio de información es un espacio abstracto consistente de un número de estados, que llamaré *estados de información*, y una estructura básica de *relaciones de diferencias* entre esos estados. El espacio de información no trivial más simple es el espacio que consiste en dos estados con una diferencia primitiva entre ellos. Podemos pensar en esos estados como los dos "bits", 0 y 1. El hecho de que estos dos estados son diferentes entre sí agota su naturaleza. Es decir, este espacio de información está completamente caracterizado por su estructura de diferencias.

Otros espacios de información son más complejos. Esto puede ocurrir de dos modos: si permitimos una estructura de diferencias más compleja entre los estados o si permitimos que los propios estados tengan estructura interna. Para ilustrar el primer modo, podríamos movernos a un espacio de cuatro estados que involucra a los estados 0, 1, 2 y 3. Para ilustrar el segundo modo, podríamos movernos a un

espacio estructurado que involucra estados como "110010101". Por supuesto, los dos modos podrían combinarse, produciendo estados doblemente complejos, como en un espacio con mensajes como "233102032". En lo que sigue, analizaré estos dos tipos de complejidad con mayor detalle.

Comencemos por el primer tipo de complejidad. Muy obviamente, existe un espacio de tres estados, un espacio de cuatro estados, etc., cuya estructura de diferencias es una extensión natural del espacio de dos estados. Por ejemplo, un elemento A, B, C o D seleccionado de un espacio de cuatro elementos contiene información en la misma clase de forma que un bit la contiene. Por supuesto, la naturaleza de los rótulos "A", "B", etc. es irrelevante acá; una vez más, lo único esencial en el espacio es su estructura.

Más importante, existen espacios de información *continuos*, cuyos estados se encuentran en un continuo análogo al de los números reales entre 0 y 1. Un espacio de este tipo tiene un número infinito de estados. Este espacio tiene una estructura de diferencias mucho más compleja que los casos previos: la estructura corresponde directamente a la topología del continuo: algunos estados están entre otros estados, algunos estados están más cercanos entre sí que otros estados, etc. Pero, como antes, podemos considerar que un solo punto seleccionado del continuo contiene información.

También podemos tener un espacio de información cuya estructura sea la de un continuo bidimensional o un continuo multidimensional, análogo a la estructura de una región del espacio n dimensional. Un solo punto seleccionado de una región de espacio tridimensional contendrá información, por ejemplo. En el caso más general, la estructura puede estar definida por la de un espacio topológico arbitrario que proporciona un conjunto con relaciones de "proximidad" o "vecindades". Los detalles de esto no importarán demasiado en lo que sigue, sin embargo, ya que sólo me ocuparé de estructuras intuitivamente familiares como la del continuo.

El segundo tipo de complejidad involucra estados con *estructura interna*. Estos estados están hechos de un número de estados más básicos que llamaré *elementos*. Un ejemplo es el espacio de estados de diez bits, análogo a "mensajes" como "1001101000". Cada estado aquí consiste de diez elementos, y cada elemento puede considerarse que pertenece a su propio *subespacio* de dos estados análogo al espacio de dos estados original. Podemos considerar este espacio de información como una especie de producto de diez subespacios, cada uno de los cuales es un espacio de información por derecho propio.

También puede haber estructuras internas más interesantes.

Por ejemplo, un estado de información podría tener una estructura interna continua, de modo que sea una especie de análogo continuo de la estructura de diez elementos que mencionamos antes. Un estado de este tipo tendría un número infinito de elementos, cada uno de los cuales corresponde a su propio subespacio. Podríamos pensar el espacio de información correspondiente como similar al espacio de funciones sobre un continuo (donde cada valor pertenece a un subespacio) o sobre un espacio continuo más complejo.

También puede ocurrir que los subespacios sean complejos en el primer modo mencionado más arriba: por ejemplo, los elementos en cada subespacio podrían caer a lo largo de un continuo. De modo que hay lugar aquí para dos niveles simultáneos de complejidad. Por ejemplo, cada estado podría consistir de una estructura continua de elementos, cada uno de los cuales puede tomar valores dentro de un subespacio continuo. Un estado de información en este espacio podría verse como una forma de onda, o como otra función con dominio y rango continuos: es un análogo continuo de los "mensajes" discretos que describimos anteriormente.

En el caso más general, un espacio de información tendrá dos tipos de estructura: cada estado complejo podría tener una estructura interna, y cada elemento en ese estado pertenecerá a un subespacio con una estructura de diferencias topológicas propia. Podríamos llamar a la primera de estas la estructura *combinatoria* del espacio y a la segunda la estructura *relacional* de los subespacios. Gran parte del tiempo, cada subespacio tendrá el mismo tipo de estructura relacional, de modo que podemos hablar simplemente de la estructura relacional del propio espacio. La estructura *general* del espacio está dada por las estructuras combinatoria y relacional juntas. Frecuentemente me limitaré a espacios de información con estructura relacional y no estructura combinatoria —el caso en el cual hay un único elemento en un estado de información— ya que la exposición es mucho más simple en este caso.

Este marco conceptual no incorpora nada que se parezca a una noción de contenido semántico, de modo que el tipo de información que aquí analizamos sólo está, en el mejor de los casos, relacionado indirectamente con la variedad semántica de información que estudian filósofos como Dretske (1981) y Barwise y Perry (1983). Podría ser posible extender el presente marco conceptual para que tenga un elemento semántico, si asociamos algún tipo de contenido semántico a cada estado de información, pero, en su estado actual, el marco es independiente de consideraciones semánticas.

Esta formalización captura la idea de Shannon de que la información involucra esencialmente un estado seleccionado entre un

número de posibilidades (en la estructura relacional de un espacio) y también captura la idea de que la información compleja puede construirse a partir de información simple (en la estructura combinatoria de un espacio). Un solo bit puede constituir información para Shannon, lo mismo que un "mensaje" largo como "10011010". Este investigador también considera el caso en que la información cae en un espacio continuo o en un espacio de funciones sobre un dominio continuo. En cada caso, lo crucial es la selección de un solo elemento dentro de un espacio de posibilidades contrastantes.

La propia concepción de Shannon suele ocuparse de la *cantidad de información* en un estado de información; esta mide cuán *específico* es un estado dentro de un espacio de información. Un estado en un espacio de información de dos estados posee un bit de información; un estado dentro de un espacio de cuatro estados posee dos bits; un estado dentro de un espacio de n elementos contiene log2 n bits. Cuando un espacio es una combinación de subespacios, un estado contiene una cantidad de información igual a la suma de las cantidades contenidas en sus elementos: así un "mensaje" de diez dígitos binarios contiene diez bits de información. Este tratamiento se aplica a espacios discretos; en los espacios continuos, la cantidad de información debe definirse más sutilmente. Aquí no me ocuparé mucho de la cantidad de información. Más bien me interesarán los propios estados de información, que podríamos pensar que están en una relación con la cantidad de información como la materia lo está con la masa.

## Información físicamente realizada

Tal como los definimos, los espacios de información son *espacios abstractos*, y los estados de información son estados abstractos. No son parte del mundo físico concreto o fenoménico. No obstante, podemos encontrar información tanto en el mundo físico como en el fenoménico, si miramos las cosas del modo apropiado. Para hacer esto, necesitamos analizar los diversos modos en los que los espacios y estados de información pueden *realizarse* en el mundo. Examinaré la realización física y la realización fenoménica en forma separada.

Parece intuitivamente evidente que los espacios y los estados de información se realizan en todas partes en el mundo físico. Podemos considerar que mi interruptor de luz realiza un espacio de información de dos estados, por ejemplo, constituido por sus estados "arriba" y "abajo". O podemos considerar que un disco compacto realiza un estado de información combinatorio consistente en una estructura compleja de bits. De la misma forma, podemos considerar que un

termostato, un libro o una línea telefónica realizan información. ¿Cómo podemos darle un sentido a esas intuiciones?

El modo natural de hacer la conexión entre los sistemas físicos y los estados de información es considerar la información realizada físicamente en términos de un eslogan formulado por Bateson (1972): la información es una *diferencia que hace una diferencia*. Aunque mi interruptor de luz puede adoptar un número infinito de posiciones en un rango continuo, la mayor parte de esa variación no hace ninguna diferencia en absoluto para mi luz. Si el interruptor está completamente para arriba, o un cuarto hacia abajo, la luz estará encendida. Por otro lado, cuando está en una posición más allá de aproximadamente un tercio hacia abajo, la luz estará apagada. En lo que a la luz respecta, sólo hay dos estados relevantes del interruptor, que podemos llamar "arriba" y "abajo". La diferencia entre estos dos estados es la única diferencia que hace una diferencia para la luz. De este modo, podemos considerar que el interruptor realiza un espacio de información de dos estados y que algunos estados físicos del interruptor corresponden a un estado de información y otros corresponden al otro.

En general, un espacio de información asociado con un objeto físico estará siempre definido respecto de un *camino causal* (en este caso, el camino desde el interruptor de luz a la luz) y un espacio de *efectos* posibles al final del camino (en este caso, el estado encendido/apagado de la luz). Los estados físicos corresponderán a los estados de información de acuerdo con sus efectos en el camino causal. Cuando dos estados físicos tienen el mismo efecto sobre el camino —como en el caso de dos posiciones del interruptor de luz que llevan a que la luz esté encendida— corresponderán al mismo estado de información. Si dividimos los estados físicos de este modo, llegaremos a un conjunto básico de diferencias físicas que hacen una diferencia; esto constituye la realización física de un espacio de información.

La estructura del espacio de información corresponderá directamente a la estructura del espacio de efectos; este será un espacio discreto o continuo. En el caso de la luz, por ejemplo, hay dos efectos relevantes en el camino causal: la luz puede estar encendida o apagada. De esta forma, puede considerarse que el interruptor realiza un espacio de información de dos estados.

Podemos tratar los espacios de información continuos de una manera similar. Si mi luz tiene un reductor de intensidad, entonces al rotar la perilla a diferentes posiciones producimos diferentes intensidades de luz en un rango continuo. (En la práctica el rango podría ser discreto, pero idealizo.) Los efectos sobre la intensidad de la luz definen un espacio de información continuo que se realiza en mi interruptor de luz. Los estados físicos del interruptor que producen la

misma intensidad de luz (estados en áreas en las que la perilla es insensible, tal vez, o estados que varían en parámetros irrelevantes tal como el color del interruptor) estarán asociados con el mismo estado de información. El espacio de estados de información tiene la estructura topológica del continuo; la estructura de diferencias entre los estados corresponde a la estructura de diferencias en el efecto sobre la intensidad lumínica.

La información realizada en un disco compacto puede también analizarse de esta manera. Un disco tiene un número infinito de estados físicos posibles, pero cuando se consideran sus efectos sobre el reproductor de discos compactos, sólo realiza un número finito de estados posibles de información. Muchos cambios en el disco —una alteración microscópica por debajo del nivel de resolución del mecanismo de lectura óptico, una pequeña rayadura en el disco o una marca grande en el lado inverso— no hacen ninguna diferencia al funcionamiento del sistema. Las únicas diferencias significativas para el estado de información del disco son aquellas que se reflejan en la salida del dispositivo de lectura óptica. Estas son las diferencias en la presencia de hoyos y superficies planas sobre el disco, que corresponden a lo que pensamos como "bits". Cualquier estado particular del disco tendrá un estado de información asociado dentro de un espacio de información grande. Los estados físicos de diferentes prensados de la misma grabación estarán asociados con el mismo estado de información, si todo sale bien. Los prensados de diferentes grabaciones o, incluso, prensados imperfectos de la misma grabación estarán asociados con diferentes estados de información, debido a sus diferentes efectos.

Este es un caso en el cual el espacio de información físicamente realizado tiene una estructura combinatoria. Cada "bit" sobre el disco compacto tiene un efecto independiente sobre el reproductor de discos, de modo que cada localización en el disco puede considerarse que realiza un subespacio independiente de dos estados. Al reunir todos esos efectos independientes, encontramos una estructura combinatoria en el espacio de efectos totales de un disco compacto, y así podemos hallar la misma estructura combinatoria en el espacio de información que el disco compacto realiza. Este espacio de información puede considerarse el producto de una gran colección de subespacios de dos estados, uno por cada hoyo o superficie plana en el disco.

Nótese que, según esta concepción, la información físicamente realizada sólo es información en la medida en que puede ser *procesada*. Como Mackay (1969) dice, "La información es lo que la información hace". Esto concuerda con el propio tratamiento de la infor-

mación de Shannon. La "cantidad de información" de Shannon mide la especificidad de un estado dentro del espacio de estados que pueden ser *transmitidos*, esto es, que pueden desempeñar papeles distintos en caminos causales diferentes (lo que Shannon llama un canal de comunicación). Para Shannon, la información es siempre un estado transmisible y la extensión de un espacio de información está implícitamente definido por la función de un transmisor. La información es una diferencia que puede hacer una diferencia en la transmisión.

Esto puede aclararse mediante el diagrama estándar de Shannon (fig. 8.1) y su comentario asociado. Una fuente de información es un conjunto de "mensajes" (estados de información), donde se supone que los diferentes mensajes son codificados mediante señales distintas por un transmisor y que señales transmitidas distintas corresponden a mensajes diferentes. Esta propiedad *define* lo que significa que los estados sean mensajes distintos. Si dos estados físicos diferentes del sistema son convertidos en la misma señal, entonces realizan el mismo mensaje. El deterioro y la pérdida de información pueden introducirse más tarde en el proceso, como lo indica el lado derecho del diagrama, pero es esencial en un canal de información que distintos estados de información produzcan diferentes efectos por medio de un transmisor. Todo esto está implícito, más que explícito, en la concepción de Shannon, en la que no existe ningún tratamiento directo de la relación entre estados físicos y estados de información. Pero, en un examen más detallado, es evidente que cuando se individualizan los estados de información, el principio de transmisibilidad realiza el trabajo.

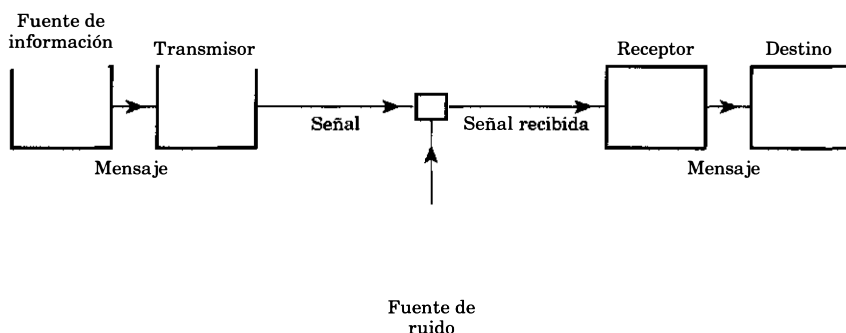



*Figura 8.1.* Diagrama de Shannon de un canal de información. (Diagrama 1, de Claude E. Shannon y Warren Weaver, *The Mathematical Theory of Communication*, 1963. Copyright 1949 por el Board of Trustees of the University of Illinois Press. Utilizado con autorización de la University of Illinois Press.)

No intentaré formular criterios precisos acerca de la realización de un espacio de información en un sistema físico. En cambio, dejaré las cosas planteadas en el nivel informal del principio de la "diferencia que hace una diferencia". Existen diversos modos en los cuales esta idea informal podría especificarse en una concepción formal, algunos de los cuales plantean restricciones más fuertes sobre la realización que otros. Sería prematuro en este punto decidirse por uno de estos en particular. Al mantener las cosas en un nivel informal, dejamos algo de espacio para maniobrar con los detalles, los que podrán clarificarse a medida que obtengamos una mejor idea de lo que es apropiado para una aplicación específica. Para el propósito de una teoría de la conciencia, especificar esos detalles de un modo correcto será parte del proceso de especificar la teoría.

## Información fenoménicamente realizada

La realización física es la manera más común de pensar acerca de la información contenida en el mundo, pero no es el único modo como puede hallarse la información. También podemos encontrar información que se realiza en nuestra *fenomenología*. Los estados de experiencia caen directamente en espacios de información de un modo natural. Existen patrones naturales de similitudes y diferencias entre estados fenoménicos, y estos patrones producen la estructura de diferencias de un espacio de información. De esta forma podemos considerar que los estados fenoménicos realizan estados de información dentro de esos espacios.

Por ejemplo, el espacio de experiencias simples del color tiene una estructura relacional tridimensional que ya hemos analizado. Si abstraemos los patrones de similitudes y diferencias entre esas experiencias, obtenemos un espacio de información abstracto con una estructura relacional tridimensional que es la que el espacio fenoménico realiza. Cualquier experiencia simple particular del color corresponde a una localización específica dentro de ese espacio. Una experiencia específica de rojo es un estado de información realizado fenoménicamente; una experiencia específica de verde es otra.

Experiencias más complejas, como las experiencias de un campo visual completo, caen dentro de los espacios de información con una estructura combinatoria compleja. Cuando miro una imagen, por ejemplo, mi experiencia cae en un espacio con (al menos) la estructura combinatoria de un continuo bidimensional, donde cada elemento en ese continuo tiene (cuanto menos) la estructura relacional tridimensional del espacio simple de colores. La estructura de los fragmentos de color en un campo visual no es *tan* diferente en tipo de la estructura

de los dígitos binarios en un mensaje de diez dígitos, aunque tanto la estructura combinatoria como la estructura relacional son mucho más complejas.

Para hallar espacios de información realizados fenoménicamente, no nos basamos en el principio de la "diferencia causal que hace una diferencia" que utilizamos para encontrar espacios de información realizados físicamente. Más bien, nos apoyamos en las cualidades intrínsecas de las experiencias y la estructura entre ellas, es decir, las relaciones de similitudes y diferencias entre ellas y su estructura combinatoria intrínseca. Cualquier experiencia tendrá relaciones naturales de similitudes y diferencias con otras experiencias, de forma que siempre podremos encontrar espacios de información en los que caen las experiencias.

## El principio del doble aspecto

Este tratamiento de la información pone de relieve un vínculo crucial entre lo físico y lo fenoménico: cada vez que encontramos un espacio de información realizado fenoménicamente, encontramos el mismo espacio de información realizado físicamente. Y cuando una experiencia realiza un estado de información, ese mismo estado de información se realiza en el sustrato físico de la experiencia.

Tómese una experiencia simple de color, que realiza un estado de información dentro de un espacio de información tridimensional. Podemos encontrar que el mismo espacio se realiza en los procesos cerebrales subyacentes a la experiencia: este es el espacio tridimensional de las representaciones neuronalmente codificadas en la corteza visual. Los elementos de este espacio tridimensional corresponden directamente a elementos del espacio de información fenoménico.

No sabemos cómo exactamente se codifican estos estados y, por lo tanto, no sabemos cómo exactamente se realiza físicamente el espacio de información. Pero sabemos que debe realizarse, ya que los procesos posteriores exhiben todos los efectos sistemáticos de la realización informacional. Nuestros informes pueden variar sistemáticamente según la localización en el espacio de color; por ejemplo, cuando evaluamos colores como relativamente "más oscuros" o "más claros". También podemos poner en correspondencia objetos con otros objetos de acuerdo con sus similitudes y diferencias de color. De manera que sabemos que debe haber diferencias significativas en la corteza visual que son *transmisibles* a otras áreas del cerebro donde producen un espacio tridimensional de posibles efectos. Los estados que subyacen a cualquier par de experiencias indistinguibles tendrán los mismos efectos, aunque podrían existir detalles físicos levemente

diferentes asociados a cada uno —piénsese en la analogía con las pequeñas diferencias en el estado del interruptor de luz—, y los estados que subyacen a dos experiencias *similares* cualesquiera tendrán efectos similares. De modo que podemos considerar que la corteza visual realiza estados de información en un espacio tridimensional.

Lo mismo ocurre para experiencias más complejas, como las de un campo visual completo. Estas se realizan en un espacio de información combinatorio, y el mismo espacio debe realizarse físicamente en los procesos cerebrales subyacentes. Sabemos que para cada localización en el campo, experiencias simples diferentes corresponden a diferencias en diversos efectos posteriores, y estas últimas diferencias son separables según su localización en el campo. Por ejemplo, podemos responder separadamente a consultas específicas acerca del color en una localización dada; este espacio separado de efectos para cada localización produce un subespacio separado para cada localización. De modo que en algún lugar en la corteza visual, debe haber una codificación de un estado de información combinatorio, para que todas las diferencias relevantes puedan ser transmitidas a los procesos posteriores. El espacio de estados posibles significativos aquí es isomórfico al espacio de experiencias posibles; de modo que podemos ver que el mismo estado de información se realiza física y fenoménicamente.

No es necesario que la información esté codificada *localmente*, por ejemplo, en una pequeña estructura de neuronas vecinas. Es posible que la información se realice físicamente de un modo holístico, como encontramos, por ejemplo, en ciertas formas holográficas de almacenamiento de información. Las diferencias significativas en los estados de la corteza visual podrían corresponder a diferencias dispersas a través de la corteza. Pero, en tanto estas sean las diferencias que son transmitidas y que tienen los efectos relevantes, la información se realizará de todos modos.

Es natural suponer que esta doble vida de los espacios de información corresponde a una dualidad en un nivel profundo. Podríamos incluso sugerir que esta doble realización es la clave de la conexión fundamental entre los procesos físicos y la experiencia consciente. Necesitamos *algún* tipo de construcción para establecer el vínculo, y la información parece una construcción tan buena como cualquiera. Podría ocurrir que los principios que conciernen a la doble realización de la información pudiesen especificarse en un sistema de leyes básicas que conecten los dominios físico y fenoménico.

Podríamos formular esta cuestión sugiriendo como principio básico que la información (en el mundo real) tiene dos aspectos, uno

físico y otro fenoménico. Allí donde hay un estado fenoménico, este realiza un estado de información, uno que también se realiza en el sistema cognitivo del cerebro. De modo recíproco, al menos para algunos espacios de información físicamente realizados, cada vez que un estado de información en ese espacio se realiza físicamente, también se realiza fenoménicamente.[1]

Este principio, por sí mismo, no se acerca a constituir una teoría psicofísica completa. Más bien, constituye una especie de plantilla para una teoría psicofísica al proporcionar un marco básico dentro del cual pueden formularse leyes detalladas. Para especificar el principio en la forma de una teoría, debe responderse todo tipo de preguntas. Por ejemplo, ¿precisamente a *cuáles* espacios de información físicamente realizados se aplica el principio básico? Analizaré esta cuestión más en detalle en el apartado 4, de modo que por ahora la dejaré sin responder. Otro tipo de subespecificidad surge de la holgura de la definición de información físicamente realizada: para una teoría psicofísica específica, necesitaremos saber precisamente cómo es que un espacio de información se realiza físicamente. Pero todo esto es parte del proceso de desarrollo de una teoría.

Algunas otras preguntas importantes conciernen a la *ontología* del punto de vista. ¿Qué tan seriamente debemos considerar el discurso del “doble aspecto”? ¿En qué medida este marco *reificará* la información, es decir, la tratará como real? ¿Afirma que lo físico, lo fenoménico o ambos son ontológicamente dependientes de lo informacional? Dejaré todas estas preguntas abiertas por ahora. Más adelante en el capítulo consideraré diversas posibles interpretaciones de la ontología. Algunas de estas interpretaciones consideran la información simplemente como una construcción útil para la caracterización de las leyes psicofísicas; otras le asignan un papel más fundamental en la ontología. De modo similar, algunas interpretaciones toman la idea de un “doble aspecto” más en serio que otras.

Por ahora, haré abstracción de estas cuestiones metafísicas. Simplemente, debe considerarse el principio como una ley que conecta los dominios físico y fenoménico, con implicaciones ontológicas que no son particularmente diferentes de las de las leyes ya consideradas. Ya sabemos que la experiencia surge de lo físico en virtud de ciertas leyes que se aplican a determinadas características físicas del mundo. La sugerencia clave aquí es que el nivel básico en el cual las leyes se aplican al mundo físico es el de la información físicamente realizada.

Por supuesto, la información puede no ser una característica *primitiva* del mundo físico, al modo como la masa y la carga lo son, pero no es necesario que sea primitiva para desempeñar un papel en las leyes psicofísicas fundamentales. Ya tenemos todas las propieda-

des fundamentales que necesitamos en las propiedades físicas y fenoménicas básicas. Lo que necesitamos ahora es una construcción que conecte ambos dominios. La información parece ser una construcción simple y directa que es apropiada para este tipo de conexión y que podría cumplir la promesa de producir un conjunto de leyes que sean simples y globales. Si puede obtenerse un conjunto de leyes de este tipo, entonces podríamos verdaderamente alcanzar una teoría fundamental de la conciencia.

Sin embargo, podría suceder que exista algún modo de considerar a la información como fundamental. La idea de que el objeto último de la física es la información ya fue examinada por algunos físicos, por ejemplo. Si esta idea llegase a tener éxito, podría ocurrir que de algún modo lo físico se derive de lo informacional y, entonces, la ontología de este enfoque podría resolverse muy pulcramente. Analizaré algunas ideas de este tipo más adelante en el capítulo.

## 3. Algunos argumentos de apoyo

No tengo argumentos contundentes que prueben que la información es la clave del vínculo entre los procesos físicos y la experiencia consciente, pero existen algunos modos indirectos de dar apoyo a la idea. Ya analizamos el primer tipo de consideración de apoyo: la observación de que los mismos espacios de información se realizan física y fenoménicamente. En lo que sigue, mencionaré otra fuente principal de apoyo y dos fuentes menores.

Las dos fuentes menores se encuentran en el hecho de que el enfoque de doble aspecto de la información es compatible con los principios psicofísicos que desarrollamos anteriormente; en particular, el principio de la coherencia estructural y el principio de la invariancia organizacional. Estos principios son fuertes restricciones y no es obvio cómo una teoría fundamental podría satisfacerlas, de modo que una característica a favor del enfoque informacional es que sea compatible con ambos.

La compatibilidad con la coherencia estructural es particularmente fácil de ver: en ciertos modos, el enfoque informacional está hecho a medida para satisfacer esta restricción. La estructura de la experiencia es exactamente la estructura de un espacio de información fenoménicamente realizado, y la estructura de la percatación es precisamente la estructura de un espacio de información físicamente realizado. Para ver el primer punto, nótese que lo que llamé la estructura implícita de una experiencia corresponde a la estructura relacional de un espacio de información y que lo que llamé la estructura explícita de una experiencia corresponde a la estructura

combinatoria del espacio. Para ver el segundo punto, nótese que los diversos detalles en la estructura de la percatación son, por definición, diferencias que hacen una diferencia en el procesamiento posterior, ya que están directamente disponibles para el control global y, por lo tanto, son la realización física de un espacio de información. A partir de que estas dos estructuras son, de hecho, realizaciones del mismo espacio de información, se deduce el principio de coherencia estructural.

Debería señalar que el principio del doble aspecto no asegura, por sí mismo, que la estructura de la percatación se proyecte en la experiencia. Para asegurarnos de ello, debemos mostrar que el espacio físico de información aquí es uno de aquellos a los que se aplica el principio del doble aspecto. Para esto, necesitaríamos una versión más detallada del principio que restrinja de un modo apropiado los espacios de información involucrados, de modo que al menos incluya información que está disponible para el control global en los casos familiares. En su estado actual, el principio del doble aspecto no *predice* todavía el principio completo de coherencia estructural, pero al menos es *compatible* con él.

Tampoco es difícil mostrar que el principio de doble aspecto es compatible con el principio de la invariancia organizacional. Para ver esto, nótese que cuando un sistema realiza un espacio de información, lo hace en virtud de su organización funcional. Cualquier otro sistema que sea funcionalmente isomórfico en un grano suficientemente fino tendrá el mismo patrón de diferencias que hace una diferencia y, por lo tanto, realizará el mismo espacio de información. De esta manera, si mis experiencias surgen en virtud de espacios de información realizados en mi cerebro, entonces los mismos espacios de información se realizarán en un isomorfo funcional y surgirán las mismas experiencias, como predice el principio de invariancia.

Una teoría fundamental de la conciencia deberá invocar características físicas que son organizacionalmente invariantes y lo bastante simples como para desempeñar un papel en las leyes fundamentales. La mayor parte de las características organizacionalmente invariantes no son muy simples y la mayor parte de las características simples no son organizacionalmente invariantes. La información físicamente realizada podría ser la característica más natural que satisface ambos criterios. El hecho de que los satisface a los dos es un punto a favor de un enfoque informacional de las leyes psicofísicas.

## La explicación de los juicios fenoménicos

Más arriba vimos que aunque la conciencia no puede explicarse reductivamente, los *juicios* fenoménicos —juicios de la forma "Yo soy consciente", "No es extraña la conciencia", etc.— pueden explicarse de ese modo, al menos en principio. Esto ponía algo de tensión sobre una teoría no reductiva de la conciencia, aunque no parecía ser fatal. Resulta contrario a la intuición que estos juicios puedan ser explicados sin invocar a la conciencia, pero es algo con lo que podemos aprender a vivir. No obstante, podemos esperar que la explicación de los juicios fenoménicos se encuentre vinculada de algún modo profundo a la explicación de la propia conciencia. Parecería irrazonable y accidental que estas dos explicaciones fuesen totalmente independientes.

Podemos formular esta cuestión como una especie de exigencia de *coherencia explicativa* sobre una teoría de la conciencia. Una teoría terminada de la mente debe proporcionar una concepción (no reductiva) de la conciencia y una concepción (reductiva) de por qué juzgamos que somos conscientes, y es razonable esperar que estas dos concepciones sean coherentes entre sí. En particular, podríamos esperar que las características de procesamiento que son centralmente responsables de generar los juicios fenoménicos serán también las que son centralmente responsables de la propia consciencia. De este modo, aunque la conciencia no sea parte de la explicación de los juicios fenoménicos, las raíces de la conciencia lo serán.

Por supuesto, no podemos *probar* que una teoría de la conciencia deba satisfacer este requerimiento, pero cualquier teoría de la conciencia que lo satisfaga tendrá un elemento de fuerza a su favor del cual otras teorías carecen. Si una teoría es capaz de mostrar cómo la explicación de los juicios fenoménicos involucra centralmente la base explicativa de la conciencia, entonces las habremos entrelazado en una imagen más unificada de la mente, y habremos podido eliminar algo de esa sensación de coincidencia escandalosa.

Frecuentemente pensé que esto podría ser la clave para encontrar una teoría de la conciencia.[2] Primero, necesitamos esforzarnos por comprender por qué se producen los juicios acerca de la conciencia. Esto podría ser una cuestión difícil, pero no debería involucrar profundos misterios metafísicos; en principio, es una cuestión en el dominio de la ciencia cognitiva. Luego, necesitamos abstraer las características clave en esa explicación y considerar cómo podrían desempeñar un papel en una teoría de la conciencia. No hay ninguna garantía de que esto nos lleve a una teoría satisfactoria, pero es una estrategia prometedora.

De cualquier forma, es probable que la búsqueda de una explicación reductiva de nuestros juicios acerca de la conciencia sea esclarecedora y una de las aplicaciones más valiosas de los métodos reductivos en el desarrollo de una teoría de la conciencia. Podríamos concentrarnos en por qué un sistema de procesamiento debería producir juicios de que está consciente y, en particular, en por qué debería juzgar que la conciencia es un fenómeno extraño. Ya dije algunas pocas palabras al respecto en el capítulo 5; aquí profundizaré la cuestión. La "explicación" que formularé sólo es una concepción aparentemente plausible, pero podemos esperar que se especificaría, con la ayuda de investigaciones empíricas, en una teoría detallada. Es probable que existan jugosas recompensas para la ciencia cognitiva y la neurociencia en abordar estos fenómenos.

De modo que dejemos de lado a la conciencia por el momento, y concentrémonos en el sistema de procesamiento cognitivo desde un punto de vista de tercera persona. Piense en esta explicación como aplicándose a un zombi, si así le gusta. ¿Por qué podríamos *esperar* que un sistema de procesamiento debería producir este tipo de juicio? ¿Qué tipo de proceso podría contribuir al juicio de que una sensación de color está presente, por ejemplo? Para pensar en ello, considere qué podría estar ocurriendo cuando percibimos colores.

Sin entrar en los detalles de bajo nivel, la historia sería aproximadamente la siguiente. Una envolvente espectral particular de luz hace impacto en nuestros ojos y activa diferentes tipos de células retinianas. Tres variedades de conos abstraen nuestra información según la cantidad de luz presente en diversos rangos superpuestos de longitudes de onda. Inmediatamente, muchas distinciones presentes en la onda lumínica original se pierden. La información se transmite a través del nervio óptico hasta la corteza visual, donde vuelve a ser transformada por el procesamiento neuronal en información que corresponde a valores en tres ejes: quizá los ejes rojo-verde, amarillo-azul y acromático. Lo que ocurre luego aún no se comprende apropiadamente, pero parece que se preserva la información que corresponde a la posición de un color particular en el espacio tridimensional, antes de ser categorizado en las categorías familiares de "rojo", "verde", "marrón", etc. Se asocian categorías verbales a esas etiquetas, y se emite un informe como "Estoy viendo rojo en este momento".

Ahora adoptemos el "punto de vista" del sistema acerca de lo que está ocurriendo. ¿Qué tipo de juicio formará? Seguramente formará un juicio como "objeto rojo allí", pero si es un sistema reflexivo y racional, también podríamos esperar que sea capaz de reflexionar sobre el propio proceso de percepción. Podríamos preguntarnos cómo la percepción "golpea" el sistema.

La característica crucial es que cuando el sistema percibe un objeto rojo, los procesos centrales no tienen acceso directo al propio objeto y tampoco tienen acceso directo a los procesos físicos que subyacen a la percepción. *A lo único* a lo que estos procesos tienen acceso es a la información del color, que es sólo una localización en un espacio de información tridimensional. Cuando se trata de informar lingüísticamente sobre la situación, el sistema no puede informar, "Este fragmento está saturado de reflejos de 500 a 600 nanometros", ya que el acceso a las longitudes de onda originales se perdió totalmente. Tampoco puede informar acerca de la estructura neuronal, "Ahora tengo una frecuencia de descarga de 50 hertz", ya que no tiene ningún acceso directo a las estructuras neuronales. El sistema sólo tiene acceso a la localización en el espacio de información.

En lo que al procesamiento central concierne, simplemente *se encuentra a sí mismo* en una posición en este espacio. El sistema es capaz de hacer distinciones, y *sabe* que es capaz de hacer distinciones, pero no tiene ninguna idea de cómo lo hace. Esperaríamos que después de algún tiempo pueda llegar a *etiquetar* las diversas localizaciones a las que es arrojado —"rojo", "verde" y similares— y de que pueda ser capaz de saber exactamente en qué estado está en cualquier momento dado. Pero cuando se le pregunta *cómo* lo sabe, no hay nada que pueda decir, además de "Sólo lo sé directamente". Si le preguntamos "¿Cuál es la diferencia entre estos estados?", no tiene respuesta más allá de "Sencillamente son diferentes" o "Este es uno de *esos*" o "Este es *rojo* y ese es *verde*". Si lo presionamos para averiguar qué significa esto, el sistema no tiene nada que decir excepto "Son diferentes cualitativamente". ¿Qué más podría decir?

Es natural suponer que un sistema que puede saber directamente la ubicación que ocupa en un espacio de información, sin tener acceso a ningún conocimiento ulterior, simplemente etiquetará los estados como primitivamente diferentes, que difieren en su "cualidad". Deberíamos esperar que estas diferencias golpeen de un modo "inmediato" al sistema: es arrojado en estos estados, los que a su vez están inmediatamente disponibles para la dirección del procesamiento posterior; no hay nada inferencial, por ejemplo, acerca de su conocimiento de en qué estado se encuentra. Mas bien, deberíamos esperar que estos estados sean relativamente "inefables": el sistema no posee acceso a ninguna información significativa ulterior, de modo que no hay nada que pueda decir acerca de los estados más allá de señalar las similitudes y diferencias entre ellos y las diversas asociaciones que podrían tener. Ciertamente, no esperaríamos que la "cualidad" fuese algo que pudiese explicitar en términos más básicos.

Podría objetarse que el sistema podría estar conformado de forma de acceder a la información como "corazonadas", de un modo similar a como podría hacerlo un sujeto con ceguera visual. Quizá podría decir, "El juicio 'rojo' sólo irrumpió en mi cabeza", sin ninguna aseveración sobre la "cualidad". Pero esto claramente sería una organización ineficiente en la que el sistema debería esperar una corazonada. ¿Qué hay de las veces —como cuando jugamos al tenis, por ejemplo— en las que necesitamos reaccionar a la información visual sin formar juicios? ¿Podemos suponer que el sistema diría, "Sencillamente me encontré sabiendo dónde estaba la pelota y haciendo lo correcto, sin *experimentarlo*"? Quizás este sea un escenario coherente, pero no parece ser un diseño natural para un sistema cognitivo. Si estuviésemos diseñando un sistema de este tipo, sería mucho más natural diseñarlo de modo que simplemente "vea" por sí mismo la diferencia entre el rojo y el verde, base su conducta de inmediato en la diferencia percibida, y responda confiada y directamente cuando se le interroga. De cualquier manera, este último es al menos *un* modo razonable en el que podríamos diseñar un sistema, y eso es todo lo que necesitamos aquí.

Dado este tipo de acceso directo a los estados de información, entonces, es natural esperar que el sistema utilice el lenguaje de la "experiencia" y la "cualidad" para describir su propio punto de vista cognitivo acerca de la percepción. No es sorprendente que todo esto le parezca bastante extraño al sistema: esos estados inefables inmediatamente conocidos, que parecen ser tan importantes para su acceso al mundo pero que son tan difíciles de precisar. Es natural suponer que esto pueda parecerle extraño al sistema, en la misma manera que la conciencia nos parece extraña a nosotros.

De modo que este es el comienzo de una explicación reductiva potencial de nuestros juicios acerca de la conciencia: estos juicios surgen debido a que nuestro sistema de procesamiento es arrojado en localizaciones del espacio de información y tiene acceso directo a esas localizaciones pero a nada más. Este conocimiento directo impresionará al sistema como una "cualidad" primitiva: sabe que los estados son diferentes, pero no puede articular esto más allá de decir "uno de *estos*". Este acceso inmediato a las diferencias primitivas lleva a juicios acerca de la misteriosa naturaleza primitiva de estas cualidades, acerca de la imposibilidad de explicitarlas en términos más básicos y a muchos de los otros juicios que con frecuencia hacemos acerca de la experiencia consciente.

En todo esto, es la información la que desempeña el papel principal. Es debido a que el sistema sólo tiene acceso a estados de

información que se forman los diversos juicios de "cualidades" primitivas. El sistema simplemente es arrojado en diferentes estados y los procesos posteriores sólo tienen acceso a la estructura de diferencias de esos estados anteriores, no a algo concreto. Lo que aquí hace el trabajo es un sistema de diferencias que hace una diferencia. Es la información, y nuestro acceso a ella, lo que explica reductivamente nuestros juicios acerca de la conciencia.

Algunos podrían concluir aquí las cosas, declarar que el misterio de la conciencia fue eliminado y que se dio una explicación. Por supuesto, no creo que esto sea correcto: sólo explicamos ciertos juicios, lo que es una cuestión mucho más simple. Pero ahora podemos utilizar el principio de coherencia explicativa para lograr alguna injerencia sobre una teoría de la conciencia. Si los estados de información que se realizan en este procesamiento son los que sustentan la principal responsabilidad de nuestros juicios acerca de la conciencia, quizás esos estados de información también tengan la responsabilidad de la propia conciencia.

De hecho, es así como yo me vi llevado en primer lugar al enfoque informacional de la conciencia. Si la base explicativa de nuestros juicios fenoménicos se encuentra en una estructura de diferencias que hacen una diferencia, es natural suponer que la base explicativa de la conciencia pudiese encontrarse en el mismo lugar. Esto explicaría por qué nuestros juicios se ajustan tan bien a los estados actuales de conciencia. Una experiencia consciente es una realización de un estado de información; un juicio fenoménico se explica mediante otra realización del mismo estado de información. En un sentido, postular un aspecto fenoménico de la información es todo lo que necesitamos para asegurarnos que esos juicios son verdaderamente correctos: *existe* un aspecto cualitativo en esa información que se muestra directamente en la fenomenología y no sólo en un sistema de juicios. Esto permite, entonces, que la conciencia sea muy coherente con la estructura cognitiva, lo que lleva a una perspectiva más estrechamente entrelazada de la mente.

Podemos también notar que existe un buen ajuste entre el papel cognitivo de los estados de información y la epistemología de la experiencia. En correspondencia con las experiencias con las cuales estoy directamente familiarizado encontramos los estados de información físicamente realizados a los que el sistema tiene acceso directo (cognitivo). El sistema forma sus juicios fenoménicos sobre la base de su acceso directo a los estados de información; esta conexión causal se vincula bastante bien con la aseveración de que la experiencia —la realización fenoménica del mismo estado de información— es

lo que *justifica* las creencias fenoménicas que se forman. En ambos lados, es el mismo estado de información el que desempeña el papel crucial; es sólo la realización física en un caso y la realización fenoménica en el otro.

Nada de esto es una prueba contundente de que el enfoque de la conciencia basado en la información deba ser correcto. Pero representa un apoyo más.

## 4. ¿Es ubicua la experiencia?

A esta altura, los lectores probablemente formarán fila para objetar que la información es *ubicua*. Encontramos información en todos lados, no sólo en los sistemas que de forma estándar consideramos que son conscientes. Mi reproductor de discos compactos realiza información; el motor de mi auto realiza información; incluso un termostato realiza información. De hecho, tal como definimos la noción, encontramos información en todo lugar en el que encontramos causalidad. Hallamos causalidad en todos lados, de modo que hallamos información en todos lados. Pero, ¿encontramos experiencia en todos lados?

Hay dos modos en los que un defensor del enfoque basado en la información podría reaccionar a esta situación. El primero y más obvio es buscar nuevas restricciones sobre el *tipo* de información que es relevante para la experiencia. No cualquier espacio de información físicamente realizado está asociado con la experiencia, sino solamente aquellos que poseen ciertas propiedades. Esto requeriría una consideración cuidadosa de cuáles podrían ser estas nuevas restricciones, y de cómo podrían encajar en las leyes fundamentales. Consideraré estrategias según estas líneas más adelante, pero por ahora deseo considerar la alternativa. Es decir, aceptar que toda información está asociada a la experiencia. Si esto es así, entonces no sólo la información es ubicua. La experiencia también lo será.

Si esto es correcto, entonces la experiencia está asociada aun a los sistemas más simples. Esta idea suele considerarse escandalosa o incluso descabellada. Pero creo que merece un detenido examen. No me resulta *obvio* que la idea esté descaminada y en ciertos modos tiene algún atractivo. Examinaré entonces las razones por las que podríamos rechazar esta perspectiva para ver si son convincentes y, simultáneamente, consideraré varias razones positivas para tomarla en serio.

## ¿Cómo es ser un termostato?

Para enfocar la imagen, consideremos un sistema de procesamiento de información que es casi máximamente simple: un termostato. Considerado como un dispositivo de procesamiento de información, un termostato sólo tiene tres estados de información (un estado lleva al enfriamiento, otro al calentamiento y otro a no tomar ninguna acción). De modo que afirmamos que a cada uno de estos estados de información le corresponde un estado fenoménico. Estos tres estados fenoménicos serán diferentes y un cambio en el estado de información llevará a un cambio en el estado fenoménico. Podríamos preguntarnos: ¿Cuál es el carácter de esos estados fenoménicos? Esto es, ¿cómo es ser un termostato?

Ciertamente, no será muy interesante ser un termostato. El procesamiento de información es tan simple que deberíamos esperar que los estados fenoménicos correspondientes sean igualmente simples. Habrá tres estados fenoménicos primitivamente diferentes y ninguna otra estructura. Quizá podamos pensar en esos estados por analogía con nuestras experiencias de negro, blanco y gris: un termostato puede tener un campo fenoménico todo negro, un campo todo blanco y un campo todo gris. Pero aun esto es imputar demasiada estructura a las experiencias del termostato, al sugerir la dimensionalidad de un campo visual y las naturalezas relativamente ricas del negro, blanco y gris. Realmente deberíamos esperar algo mucho más simple, para lo que no tenemos ningún análogo en nuestra experiencia. Es probable que nos resulte tan difícil imaginar estas experiencias como a una persona ciega imaginarse cómo es ver o a un ser humano imaginarse cómo es ser un murciélago; pero, al menos, podemos intelectualmente saber algo acerca de su estructura básica.

Para hacer que la perspectiva parezca menos descabellada, podemos pensar en qué podría ocurrirle a la experiencia cuando nos movemos hacia abajo en la escala de complejidad. Comenzamos con los casos familiares de los seres humanos, en los que un procesamiento de información muy complicado da origen a nuestras familiares y complejas experiencias. Si nos movemos hacia sistemas menos complejos, no parece haber muchas razones para dudar de que los perros sean conscientes, o aun los ratones. Algunas personas lo pusieron en duda, pero creo que esto se debe frecuentemente a una confusión de la conciencia fenoménica con la autoconciencia. Los ratones podrían no tener un gran sentido de sí mismos y podrían no ser propensos a la introspección, pero parece totalmente plausible que hay *algo*

que es como ser un ratón. Los ratones perciben su ambiente por medio de patrones de flujos de información de un modo no muy diferente al de nuestro propio cerebro, aunque considerablemente menos complejo. La hipótesis natural es que en correspondencia con la "variedad perceptual" del ratón, que sabemos que tienen, existe una "variedad fenoménica". La variedad perceptual del ratón es bastante rica —un ratón puede hacer muchas distinciones perceptuales—, de modo que su variedad fenoménica también podría ser bastante rica. Por ejemplo, es plausible que a cada distinción que el sistema visual del ratón es capaz de hacer y utilizar para percibir el ambiente le corresponda una distinción fenoménica. No podemos *probar* que esto sea así, pero pareciera que es el modo más natural de pensar acerca de la fenomenología de un ratón.

Si nos movemos hacia abajo en la escala, a través de los lagartos y peces a las babosas, podemos aplicar consideraciones similares. No parece haber muchas razones para suponer que la fenomenología podría extinguirse mientras persistiría una psicología perceptual razonablemente compleja. Si así ocurriese, entonces existiría una discontinuidad radical entre las experiencias complejas y ninguna experiencia o en algún lugar en la línea la fenomenología comenzaría a salir de sincronía con la percepción, de modo que durante un trecho existiría una variedad perceptual relativamente rica acompañada por una variedad fenoménica mucho más empobrecida. La primer hipótesis parece improbable y la segunda sugiere que los sistemas intermedios tendrían vidas interiores extrañamente disociadas de sus capacidades cognitivas. La alternativa es por lo menos no menos plausible. Podemos suponer que es mucho menos interesante ser un pez que ser un humano; una fenomenología más simple estaría en correspondencia con su psicología más simple, pero parece bastante razonable suponer que hay *algo* allí.

A medida que nos movemos hacia abajo en la escala desde los peces y las babosas, a través de redes neuronales simples, hasta los termostatos, ¿dónde debería extinguirse la conciencia? Es probable que la fenomenología de los peces y las babosas no sea primitiva sino relativamente compleja, y refleje las diversas distinciones que estos seres pueden hacer. Antes de que la fenomenología desaparezca del todo, podemos suponer que llegaremos a algún tipo de fenomenología máximamente simple. Me parece que el lugar más natural para que esto ocurra es un sistema con una "psicología perceptual" correspondientemente simple, como un termostato. El termostato parece realizar el tipo de procesamiento de información de un pez o una babosa reducido a su forma más simple, de modo que tal vez podría tener el tipo correspondiente de fenomenología en su forma

más despojada. Hace una o dos distinciones pertinentes de las que depende la acción; a mí, por lo menos, no me parece irrazonable que pudiese haber distinciones asociadas en la experiencia.

Por supuesto, existen otros modos como las cosas *podrían* suceder cuando nos movemos hacia abajo en la escala de complejidad, y esto no es ningún tipo de demostración de que los termostatos *deban* tener experiencias. Pero este parece ser un modo razonable de que las cosas ocurran y, si se reflexiona sobre ello, tal vez un modo tan natural como cualquiera. Podría argumentarse que el razonamiento aquí involucrado es sólo una extensión del razonamiento por el cual atribuimos experiencia a los perros o a los ratones. Al menos, una vez que comenzamos a pensar acerca de qué podría estar ocurriendo en la experiencia de un ratón y su base en su psicología perceptual, la extensión a sistemas más simples comienza a parecer mucho más natural de lo que habría parecido al principio.

Alguien que encuentre que es "descabellado" suponer que un termostato pueda tener experiencias al menos nos debe una explicación de *por qué* lo es. Podría suponer que esto se debe a que existe una propiedad que el termostato no posee y que obviamente se requiere para la experiencia; pero a mí, al menos, ninguna propiedad de este tipo me resulta obvia. Quizás exista un ingrediente crucial en el procesamiento del que el termostato carece y que un ratón posee, o del que un ratón carece y un ser humano posee, pero no veo ningún ingrediente de este tipo que sea *obviamente* requerido para la experiencia, y ciertamente no es evidente que deba existir.

Por supuesto, decir que los termostatos tienen experiencia no significa decir que tengan mucho en lo que se refiere a una vida mental. Un termostato no será *auto*consciente; no será inteligente en lo más mínimo; y yo no afirmaría que un termostato puede *pensar*.[3] Algo de la resistencia a la idea de un termostato consciente podría provenir de confundir la experiencia con estas otras características mentales, todas las cuales casi seguramente requieren mucha mayor complejidad. Estas características tienen un gran componente psicológico, y es probable que se necesite un sistema complejo para soportar los papeles causales relevantes. Pero, si distinguimos las propiedades fenoménicas de las propiedades psicológicas, la idea de un termostato consciente parece menos amenazante. Sólo necesitamos imaginar algo así como un "destello" inarticulado de experiencia, sin conceptos, sin pensamiento o cualquier otro procesamiento complejo en la vecindad.

Otra razón de por qué algunos podrían rechazar la idea de un termostato consciente es que no podemos encontrar ningún *espacio* para la conciencia en el sistema. En apariencia es demasiado simple

y no parece que la conciencia pueda tener ningún papel. Sin embargo, tener esta reacción es haber fracasado en aprender la lección del enfoque no reductivo: nunca encontraremos la conciencia dentro de un sistema aunque hagamos un examen detallado y siempre podremos comprender el procesamiento sin invocar la conciencia. Si esta no es lógicamente superveniente, no deberíamos esperar tener que encontrar "espacio" para la conciencia en la organización de un sistema; esta es bastante distinta de las propiedades de procesamiento del sistema.

Podría ocurrir que algunos no estén dispuestos a aceptar la posibilidad de termostatos conscientes simplemente porque *comprendemos* estos artefactos demasiado bien. Conocemos todo acerca de su procesamiento y no parece haber ninguna razón para invocar la conciencia. Pero, en este aspecto, los termostatos no son realmente diferentes del cerebro. Una vez que hayamos comprendido perfectamente el procesamiento cerebral, tampoco tendremos ninguna razón para invocar la conciencia. La única diferencia es que, en este momento, lo que ocurre dentro de un cerebro es lo suficientemente misterioso como para que nos sintamos tentados a suponer que la conciencia está de algún modo "localizada" en aquellos procesos cerebrales que aún no comprendemos. Sin embargo, como argumenté antes, aun cuando lleguemos a comprender estos procesos, ello no bastará para incorporar la conciencia al cuadro; de modo que aquí, una vez más, cerebros y termostatos están a la par.

Podría molestarnos el hecho de que podemos *construir* un termostato sin colocar en él ninguna conciencia. Pero, por supuesto, lo mismo se aplica a un cerebro, al menos en principio. Cuando construimos un cerebro (en la reproducción o en el desarrollo, digamos), la conciencia es un acompañamiento gratuito; lo mismo ocurrirá con un termostato. ¡No deberíamos esperar localizar la conciencia como un componente físico del sistema! Algunos podrían preocuparse por el hecho de que un termostato no está *vivo*; pero es difícil ver por qué eso debería constituir una diferencia de principio. Podemos suponer que un cerebro de silicio aislado, del tipo que consideramos en el último capítulo, no calificaría como viviente, pero hemos visto que podría ser consciente. Si los argumentos en el último capítulo son correctos, entonces el hecho de que un termostato no está compuesto de elementos *biológicos* no hace, en principio, ninguna diferencia.

Algo de la resistencia intuitiva podría provenir del hecho de que no parece haber espacio en un termostato para alguien o algo que *tenga* experiencias: ¿dónde en un termostato podría caber un *sujeto*? Pero no deberíamos buscar un homunculus en los sistemas físicos que sirva de sujeto. El sujeto es todo el sistema o, mejor, está asociado con

el sistema al modo como un sujeto está asociado con un cerebro. La forma correcta de hablar acerca de esto es difícil. Estrictamente, no diríamos que mi *cerebro* tiene experiencias, sino que *yo* tengo experiencias. Como sea que comprendamos esta relación, lo mismo se aplicará a los termostatos: estrictamente hablando, es probable que sea mejor no decir que el termostato tiene las experiencias (aunque continuaré haciéndolo cuando hablo informalmente), sino que las experiencias están asociadas con el termostato. No encontraremos a un sujeto "dentro" del termostato, así como no lo encontramos dentro de un cerebro.

Pero, volvamos a los puntos positivos en favor de sistemas simples que tienen experiencias. De este modo, evitamos la necesidad de que la conciencia "aparezca" en un cierto nivel de complejidad. Hay algo extraño en la idea de que un sistema con $n$ elementos no pueda ser consciente, pero un sistema con $n + 1$ elementos pueda serlo. Además, no podemos evitar tomar una decisión al modo como podríamos evitar tomarla acerca de cuándo precisamente alguien se vuelve "calvo": en este último caso, es plausible que exista un grado de indeterminación semántica, pero es mucho menos plausible que pueda dejarse indeterminado si un sistema es consciente. (Esto es válido especialmente si adoptamos un enfoque no reductivo, según el cual no podemos explicitar los hechos acerca de la experiencia en términos de hechos más básicos, así como explicitamos cuestiones indeterminadas acerca de la calvicie en términos de hechos determinados acerca del número de pelos sobre una cabeza.) Aunque *podría* ocurrir que la experiencia aparezca en un punto particular, cualquier punto específico parece arbitrario, de modo que una teoría que evite tener que tomar esta decisión consigue una cierta simplicidad.

Una consideración final en favor de sistemas simples que tienen experiencia: si esta es en verdad una propiedad fundamental, parece que sería natural que estuviese ampliamente difundida. Todas las otras propiedades fundamentales que conocemos ocurren incluso en sistemas simples y en todo el universo. Sería extraño que una propiedad fundamental se instancie por primera vez sólo en un momento relativamente tardío de la historia del universo e incluso entonces sólo en sistemas complejos ocasionales. No hay ninguna *contradicción* en la idea de que una propiedad fundamental sólo se instancie ocasionalmente; pero la alternativa parece más plausible, si todo el resto permanece igual.

## ¿Hacia dónde va el panpsiquismo?

Si existe la experiencia asociada a los termostatos, probablemente exista la experiencia *en todos lados*: dondequiera que haya una interacción causal, hay información, y dondequiera que haya información, hay experiencia. Podemos encontrar estados de información en una piedra —cuando se expande y contrae, por ejemplo— o incluso en los diferentes estados de un electrón. De modo que si el principio no restringido del doble aspecto es correcto, habrá experiencia asociada con una piedra o un electrón.

(No diría exactamente que una piedra *tiene experiencias* o que una piedra *es consciente*, al modo como podría decir informalmente que un termostato tiene experiencias o es consciente. Una piedra, a diferencia de un termostato, no se la considera un sistema de procesamiento de información. Simplemente se la ve como un objeto, de modo que la conexión con la experiencia es menos directa. Sería mejor decir que una piedra *contiene* sistemas que son conscientes: es posible que haya muchos de esos subsistemas, ningunas de cuyas experiencias puede considerarse canónicamente como las de la piedra [no más de lo que mis experiencias pueden considerarse las de mi oficina]. Para el termostato, en cambio, existe un espacio de información canónico asociado, de modo que parece más razonable hablar de las experiencias canónicas del termostato. Por supuesto aun este uso es algo informal, como se hizo notar más arriba.)

El enfoque de que hay experiencia dondequiera que haya interacción causal es contrario a la intuición. Pero es un punto de vista que puede resultar sorprendentemente satisfactorio si se reflexiona sobre él, ya que hace que la conciencia se integre mejor en el orden natural. Si el enfoque es correcto, la conciencia no ocurre en repentinos destellos, con sistemas complejos aislados produciendo arbitrariamente experiencias conscientes ricas. En cambio, es una propiedad más uniforme del universo, donde los sistemas muy simples tienen una fenomenología muy simple y los sistemas complejos tienen una fenomenología compleja. Esto hace que la conciencia sea menos "especial" en algunos modos y, de esta manera, más razonable.

Una pregunta interesante es si se requiere causalidad *activa* para la experiencia. ¿Podría un termostato tener experiencia cuando permanece en un estado constante (en un sentido, "causando" una salida, pero sin realmente *hacer* nada)? ¿O tiene experiencia sólo cuando está en un estado de flujo? La mayor parte de la causalidad que subyace en la experiencia en el cerebro parece ser activa, en el sentido de que la información significativa está siendo procesada constantemente, las neuronas descargan, etc. Por otro lado, podría

ocurrir que en un nivel fundamental no pueda hacerse la distinción entre causalidad activa y pasiva, en cuyo caso las dos podrían ser tratadas en un pie de igualdad. No conozco la respuesta de esta pregunta, pero pareciera intuitivamente que se requiere algún tipo de actividad para la experiencia.

Una posibilidad que no consideré hasta ahora pero que no puede desecharse es que los sistemas simples no tengan propiedades fenoménicas, pero que tengan propiedades *protofenoménicas*. Mencioné en el capítulo 4 la posibilidad de que podrían existir propiedades más fundamentales que las propiedades fenoménicas de las que estas últimas estarían constituidas. Si realmente existiesen estas propiedades, entonces parecería natural que estuviesen instanciadas en sistemas simples. Si esto es así, entonces, los termostatos podrían no tener experiencias tal como nosotros usualmente pensamos en ellas pero, en cambio, instanciarían un tipo relacionado de propiedad que no comprendemos totalmente (un tipo de protoexperiencia, tal vez). Esto permitiría conservar el enfoque unificado del orden natural mencionado más arriba, y podría también ayudar con el problema de la "extinción" (si las propiedades protofenoménicas son fundamentales, entonces las experiencias constituidas por estas propiedades podrían "aparecer" gradualmente después de todo). Dado que no sostiene que los termostatos tengan experiencias completas, este enfoque también parece un poco menos "descabellado" que la alternativa. Por supuesto, el costo es la postulación de una clase de propiedades no familiares que no comprendemos; pero la posibilidad debe dejarse abierta.

De cualquier modo, este enfoque tiene mucho en común con lo que se conoce frecuentemente como *panpsiquismo*, el enfoque de que todo tiene una mente. Existen algunas razones por las que yo no suelo utilizar este término: 1) porque creo que tener experiencias podría no ser suficiente para lo que usualmente se piensa que es tener una mente, aunque podría considerarse como una mente en su forma más simple; 2) porque las propiedades protofenoménicas podrían estar aun más lejos del concepto usual de "mente", 3) porque no pienso que sea estrictamente exacto decir que las piedras (por ejemplo) tienen experiencias, por las razones mencionadas arriba, aunque las piedras pueden tener experiencias asociadas con ellas. Tal vez la razón principal de que el término sea equívoco, sin embargo, es que sugiere una perspectiva según la cual las experiencias en los sistemas simples como los átomos son fundamentales y las experiencias complejas serían de alguna manera la suma de dichas experiencias más simples. Aunque este es un modo en el que las cosas podrían suceder, no hay ninguna razón de que deban ser así: las experiencias complejas

pueden ser más autónomas de lo que esto sugiere. En particular, el enfoque informacional sugiere un cuadro en la cual las experiencias complejas están determinadas más holísticamente.

Habiendo hecho estas advertencias, probablemente sea justo decir que este enfoque es una variedad del panpsiquismo. Debemos notar, sin embargo, que el panpsiquismo no está en la base metafísica de mi enfoque: lo que está en la base es más bien el dualismo naturalista con leyes psicofísicas. El panpsiquismo es meramente una forma en que podría funcionar la superveniencia natural de la experiencia a lo físico. En cierto sentido, la superveniencia natural proporciona el *marco conceptual*; el panpsiquismo es tan sólo un modo de resolver los *detalles*.

En lo personal, tengo mucha más confianza en el dualismo naturalista que en el panpsiquismo. Esta última cuestión parece mucho más abierta. Pero espero haber dicho lo suficiente para mostrar que deberíamos tomar en serio la posibilidad de algún tipo de panpsiquismo: no parece existir ningún argumento concluyente en contra de ese enfoque y existen varias razones positivas para adoptarlo.

## La restricción del principio de doble aspecto

Aunque estemos preparados para aceptar que los sistemas muy simples tienen experiencias, la idea de que toda información está asociada a la experiencia podría todavía hacernos sentir incómodos. Por ejemplo: sólo una pequeña cantidad de la información en el procesamiento cognitivo humano parece corresponder a la información en la experiencia consciente. ¿No es un hecho que la mayor parte de nuestro procesamiento de información es inconsciente?

Si el principio no restringido del doble aspecto es correcto, entonces podemos suponer que la respuesta es que toda esa información "inconsciente" se realiza en la experiencia; sólo que no se realiza en *mi* experiencia. Por ejemplo, si existe experiencia asociada con una de mis neuronas, al modo como existe experiencia asociada con un termostato, no esperaríamos que esta sea parte de *mi* experiencia, así como no esperaríamos que mi experiencia se transforme radicalmente si esa neurona fuese reemplazada por un pequeño homunculus consciente. De modo similar, podría haber experiencia asociada a diversos subsistemas "inconscientes" de procesamiento de información en el cerebro: es sólo que esas experiencias pertenecen a un sujeto diferente. Existen muchos sistemas diferentes de procesamiento de información en el cerebro y el que corresponde a *mí* —quizás el sistema que hace que alguna información esté disponible para un cierto tipo de control global y comunicación— es

sólo uno de ellos. Yo no esperaría tener acceso directo a las experiencias de los otros sistemas, así como tampoco esperaría tener acceso directo a las experiencias de otros seres humanos.

También podríamos preocuparnos acerca de todos los sistemas de procesamiento de información relativamente complejos en el mundo, que se encuentran en cualquier lugar desde mi reproductor de discos compactos hasta mi estómago. ¿Todos ellos califican como individuos conscientes con experiencias complejas? Para dar una respuesta, es importante observar que estos sistemas no tienen nada que se parezca a la estructura cognitiva coherente de nuestro propio sistema, de modo que es probable que cualquier experiencia asociada no se parezca en nada a las nuestras. Si un reproductor de discos compactos tiene asociadas experiencias, por ejemplo, es probable que no sean más que una estructura "plana" de bits; y si la información en mi estómago está asociada con experiencias, entonces no hay ninguna razón para pensar que estas correspondan al tipo de cosas que pensamos como una mente. Los tipos de experiencias que nosotros tenemos únicamente surgirán cuando los sistemas de procesamiento de información hayan sido formados por la evolución para tener estructuras cognitivas complejas y coherentes que reflejen una rica representación del mundo externo. Es probable que sólo un grupo muy restringido de sujetos de experiencia tengan la estructura psicológica requerida para verdaderamente calificar como *agentes* o como *personas.*

No obstante, esta gran proliferación de las experiencias, en especial la proliferación dentro de un solo cerebro, podría ser una causa de incomodidad. Esto se exacerba si se advierte que cuando tenemos un espacio de información, por lo general es fácil encontrar muchos espacios de información levemente distintos si simplemente individualizamos un camino causal significativo de un modo diferente o si dividimos los efectos significativos (las "diferencias" que la información hace) de una manera levemente distinta. ¿Debemos suponer que existen conjuntos diferentes de experiencias para todos esos espacios de información? Si esto es así, entonces ¡yo podría tener un número de parientes fenoménicos muy cercanos pero levemente diferentes que surgirían de los procesos que ocurren en mi propia cabeza!

La alternativa es *constreñir* el principio del doble aspecto de modo de reducir la clase de espacios de información físicamente realizados que tienen contrapartes fenoménicas. La estrategia más natural podría ser restringir el *modo* como se procesa la información. Después de todo, ya dije que la información en mi sistema que corresponde más directamente a mi experiencia es la información que está directamente disponible para el control global. En su estado

actual, es muy improbable que este "criterio" tenga un papel en una ley fundamental, ya que es una noción demasiado vaga y de alto nivel; podemos utilizar el principio únicamente si ya hemos individualizado un sistema de alto nivel como una persona o un cerebro. Pero tal vez exista un criterio más simple y preciso que pueda hacer el trabajo.

Una posibilidad es que la *amplificación* de la información sea crucial. La información físicamente realizada se realiza también en la experiencia sólo si la información se *amplifica* de ciertos modos, volviéndose disponible para hacer una gran diferencia a lo largo de ciertos caminos causales. Tal vez incluso podríamos decir que la intensidad de una experiencia corresponde al grado de amplificación, o algo así. Esto podría concordar bien con el criterio de disponibilidad global, aunque podría tener otros problemas: hay mucha información amplificada que no es intuitivamente consciente, por ejemplo; tampoco es obvio cómo debería precisarse la noción de amplificación.

Otra posibilidad es que podríamos restringir el tipo de *causalidad* involucrada en un sistema. Hemos visto que dondequiera que haya causalidad, hay información; pero, quizá, sólo un cierto tipo de causalidad cuenta en la individualización de los espacios de información que subyacen en la experiencia. Tal vez sólo sean significativos ciertos tipos de relaciones causales "activas", por ejemplo, o tal vez se requieran ciertos tipos de relaciones causales "naturales". Intuitivamente podría ocurrir que muchos de los espacios de información que pueden hallarse mediante los criterios dados hasta ahora sean en cierto sentido no naturales; quizás exista un modo de clarificar la restricción pertinente. Es probable que esto todavía permita una clase muy amplia de estados de información, pero podría impedir su proliferación astronómica.

No estoy seguro de cuál debería ser el criterio de restricción relevante, pero esto no significa que no pueda haber alguno. Podría ocurrir incluso que un criterio de restricción constriña los espacios de información relevantes de manera que la información en sistemas simples como los termostatos no califiquen. Mi propia intuición me dice que bien podría existir una restricción sobre el principio del doble aspecto, pero que la información en sistemas simples como los termostatos podrían calificar de todas formas. Por mi parte, la proliferación de muchas experiencias relacionadas en el cerebro me parece menos intuitiva que la presencia de experiencias en sistemas simples, aunque ninguna de las dos cuestiones está claramente definida. De cualquier forma, hay muchos modos diferentes en los que las cosas podrían ocurrir, cuando la prototeoría se elabora en una teoría.

## 5. La metafísica de la información

La pregunta sigue vigente: ¿Cómo entendemos la ontología del enfoque del doble aspecto de la información? ¿Qué seriedad le asignamos al discurso de los espacios y los estados de información: son estos solamente construcciones útiles o son de algún modo ontológicamente fundamentales? ¿Es primordial la información, o en realidad son lo físico y lo fenoménico los que son primordiales, y la información sólo proporciona un vínculo útil?

Existen varias maneras de entender esto. La más directa y menos aventurada es considerar que las realizaciones física y fenoménica de la información son características totalmente independientes, sin ningún vínculo ontológico más allá de una conexión legaliforme y un tipo de isomorfismo estructural. Según este enfoque, la ontología sigue siendo la ontología del dualismo de propiedades, con propiedades físicas, propiedades fenoménicas separadas, y una conexión legaliforme entre las dos. Aquí, la expresión "doble aspecto" debe interpretarse de un modo deflacionario: es meramente un modo colorido de hablar acerca de dos tipos diferentes de propiedades correlacionadas con una estructura similar. La información es simplemente una herramienta útil para caracterizar esta estructura común; no corresponde a ninguna categoría ontológica "profunda".

Este podría ser un modo perfectamente adecuado de ver las cosas, pero existen algunas posibilidades más interesantes. La mayoría de ellas involucran tomar más en serio el papel de la información. Consideraré una forma de hacer esto en lo que sigue. Debo advertir al lector que este análisis pertenece al dominio de la metafísica especulativa, pero probablemente esta sea inevitable para poder llegar a un acuerdo con la ontología de la conciencia.

### *It de bit*

A veces se sugiere dentro de la física que la información es fundamental para la física del universo, e incluso que las propiedades y leyes físicas podrían derivarse de propiedades y leyes de la información. Esta perspectiva, denominada "it de bit", fue formulada por Wheeler (1989, 1990) y Fredkin (1990), y se investiga también en artículos en Zurek (1990) y Matzke (1992, 1994). Si esto es así, entonces podríamos darle a la información un papel más serio en nuestra ontología. Para tener una mejor comprensión de la cuestión, analizaré un modo clave en el cual la información puede considerarse fundamental en la física. Esta no es la única manera como las ideas de "it de bit" podrían formularse (en particular difiere en cierto modo

del punto de vista de Wheeler),[4] pero me parece que tal vez sea el modo más natural de comprender la noción. Esta interpretación está estrechamente relacionada, como veremos, con las ideas "russellianas" que analizamos en el capítulo 4 (pp. 203-05).

Esta perspectiva surge de la observación de que en las teorías físicas los estados físicos fundamentales son efectivamente individualizados como *estados de información*. Cuando examinamos una característica como la masa o la carga, simplemente encontramos un espacio primitivo de diferencias que hacen una diferencia. La física no nos dice nada acerca de qué *es* la masa o qué *es* la carga: simplemente nos dice el rango de valores diferentes que estas características pueden tomar y nos dice cuáles son sus efectos sobre otras características. En lo que a las teorías físicas concierne, los estados específicos de masa o carga podrían ser puros estados de información: todo lo que importa es su localización dentro de un espacio de información.

Esto se refleja en el hecho de que la física no se compromete en absoluto con el modo como esos estados se *realizan*. Cualquier realización de los estados de información servirá igualmente bien para los propósitos de una teoría física, siempre que mantenga la estructura correcta de relaciones causales o dinámicas entre estados. Después de todo, en tanto la forma de estas relaciones sea la misma, la física *parecerá* la misma para nuestros sistemas perceptuales: no tenemos acceso a ninguna propiedad ulterior de la realización en el mundo externo, más allá de la forma de la red causal. (Excepto, tal vez, en la medida que nuestras propiedades fenoménicas estén ligadas directamente a propiedades realizadoras.)

A veces se insinuó, incluso, que el universo podría ser un gigantesco ordenador. Fredkin (1990) sugirió que el universo podría ser un inmenso autómata celular, realizado en el fondo en una vasta estructura de bits.[5] Leckey (1993) sugirió que todo el espacio y el tiempo podría estar basado en un proceso computacional, con registros separados para cada característica fundamental instanciada del mundo. En tanto estos registros tengan las relaciones causales apropiadas entre ellos, ninguna de las criaturas en ese mundo será más sabia. El ejemplo del ordenador ilustra la amplia variedad de modos como las entidades físicas que "conocemos" podrían realizarse, siempre que haya entidades que desempeñen los papeles causales apropiados. Esto podría considerarse como parte de la "metafísica de la física": la especulación acerca de la ontología subyacente a la estructura causal del propio espacio y tiempo.

Este tipo de metafísica es claramente algo de lo que la propia física no se ocupa. La física puede permanecer neutral acerca de estas

cuestiones de cómo se realizan sus características y si las características se "realizan" de ese modo. En lo que a la física concierne, el estado del mundo podría igualmente *agotarse* en una caracterización de la información. Si existen otras propiedades "realizadoras" subyacentes ulteriores, estas no tienen ningún papel directo en las teorías físicas. De manera que podríamos sentir la tentación de directamente desecharlas.

Esto llevaría a una imagen del mundo como un mundo de *pura* información. A cada característica del mundo le corresponde un espacio de información, y donde sea que la física considere que esas características están instanciadas, un estado de información del espacio relevante es instanciado. En tanto los estados de información tengan las relaciones correctas entre ellos, todo será como debe ser. Según esta imagen del mundo, no hay nada más que decir. La información es todo lo que hay.

Así es como yo interpreto la concepción "it del bit" del mundo. Es una concepción extrañamente hermosa: una imagen del mundo como puro flujo de información, sin ninguna sustancia ulterior en él. (Algunas versiones de este enfoque pueden también aceptar el *espacio y tiempo* como un marco primitivo dentro del cual los espacios de información están inmersos; otras versiones consideran que el propio espacio y tiempo está constituido por las relaciones entre los espacios de información.) El mundo es simplemente un mundo de diferencias primitivas y de relaciones causales y dinámicas entre esas diferencias. Según este enfoque, tratar de decir alguna otra cosa acerca del mundo es un error.

## La fundamentación de la información en la fenomenología

Parece haber dos problemas principales con esa imagen del mundo. El primero lo plantea la propia conciencia. Parece que aquí tenemos algo que va más allá de un *puro* espacio de información. Las propiedades fenoménicas tienen una naturaleza intrínseca, una que no se agota en su localización en un espacio de información, y parece que un enfoque puramente informacional del mundo no deja espacio para esas cualidades intrínsecas.

El segundo problema es que no es obvio que la noción de puro flujo de información sea coherente. Podríamos creer que según este enfoque el mundo es demasiado carente de sustancia para *ser* un mundo. ¿Podría haber diferencias que sean diferencias *primitivas*, no basadas en diferencias en ninguna cualidad subyacente? Podríamos encontrar que es plausible que toda diferencia concreta en el mundo deba tener una base: esto es, que deba haber una diferencia *en* algo.

Este problema está estrechamente relacionado con los problemas de la perspectiva del “puro flujo causal” que analizamos en el capítulo 4 (p. 203), del que este enfoque es una variante. Ese punto de vista eliminaba del mundo todas las cualidades intrínsecas y dejaba un mundo de relaciones causales sin nada, parecía, que realizase esa causalidad. El presente enfoque podría ser algo mejor ya que permite estados de información como aquello que las relaciones causales relacionan, pero esos estados son notablemente insustanciales, ya que sólo son *diferentes* entre sí y no tienen ninguna naturaleza propia. Podríamos pensar que esta imagen de un mundo sin naturaleza intrínseca no es una imagen de un mundo en absoluto.

Podríamos argumentar que los espacios de información *deben* tener algo de naturaleza ulterior. Podría ocurrir que dos propiedades fundamentales tengan el mismo tipo de estructura de información; por ejemplo, que ambas involucren cantidades reales sobre un continuo. Si la física es pura información, no habrá nada que distinga las instanciaciones de los dos espacios de información. Pero debe haber *alguna* diferencia entre ellos, ya que las dos propiedades participan en leyes distintas y tienen efectos diferentes en otras características del mundo. De manera que debe haber algo ulterior que distinga esas instanciaciones; algo que vaya más allá de la pura información. Parecería que es necesario algún tipo de cualidad intrínseca para hacer la distinción.

Existen algunos modos en los que podríamos tratar de solucionar estos problemas. Podríamos decidir que el segundo problema no es, al fin de cuentas, un problema fatal y contentarnos con una física de la pura información; luego podríamos intentar incorporar propiedades fenoménicas de alguna manera vinculadas de un modo legaliforme con esa información. Alternativamente, podríamos responder el segundo problema postulando propiedades intrínsecas en las que estén basados los espacios de información físicos, y resolver el primer problema introduciendo propiedades fenoménicas de un modo separado.

La estrategia más interesante, sin embargo, es intentar responder los dos problemas a la vez. El primer problema sugiere que tenemos conocimiento directo de alguna naturaleza intrínseca en el mundo, más allá de la pura información, en las propiedades fenoménicas; y el segundo sugiere que podemos *necesitar* alguna naturaleza intrínseca en el mundo sobre la cual basar los estados de información. Tal vez, entonces, la naturaleza intrínseca requerida para basar los estados de información esté estrechamente relacionada con la naturaleza intrínseca presente en la fenomenología. Tal vez, incluso, una sea constitutiva de la otra. De este modo, obtenemos una

ontología económica y elegante, y resolvemos dos problemas de un solo golpe.

Una vez más, esto está estrechamente vinculado a la sugerencia russelliana descripta en el capítulo 4, según la cual las propiedades intrínsecas desconocidas del mundo son consideradas propiedades fenoménicas (o protofenoménicas). Russell necesitaba que esas propiedades subyacieran a las relaciones causales propuestas por la física y nosotros las necesitamos para basar los estados de información (las diferencias que hacen una diferencia) postuladas por la física. Estos son esencialmente un mismo problema. En los dos casos, tenemos la sensación de dos soluciones por el precio de una. Necesitamos que algunas propiedades intrínsecas le den un sentido al mundo físico y necesitamos encontrar un lugar para las propiedades intrínsecas reveladas en la fenomenología. Los dos problemas parecen tener una buena correspondencia.

De esta manera, la sugerencia es que los espacios de información requeridos por la física están basados en propiedades fenoménicas o protofenoménicas. Cada instanciación de un espacio de información de este tipo es, de hecho, una realización *fenoménica* (o protofenoménica). Cada vez que una característica como la masa y la carga se realizan, existe una propiedad intrínseca detrás de ella: una propiedad fenoménica o protofenoménica, o una propiedad *microfenoménica* para abreviar. Tendremos un conjunto de espacios microfenoménicos básicos, uno para cada propiedad física fundamental, y son estos espacios los que basarán los espacios de información que la física requiere. Las diferencias últimas son estas diferencias microfenoménicas.

Por supuesto, este punto de vista requiere nuevamente una variedad de panpsiquismo "escandaloso", pero ya argumenté que un panpsiquismo de este tipo no es tan irrazonable como comúnmente se supone. Con anterioridad sugerí que puede haber propiedades fenoménicas dondequiera que haya información, así que bien podemos ponerlas a desempeñar un papel útil.

La ontología a la que esto nos lleva podría verdaderamente llamarse una ontología de doble aspecto. La física requiere estados de información pero sólo le importan sus relaciones, no su naturaleza intrínseca; la fenomenología requiere estados de información, pero sólo le importa su naturaleza intrínseca. Este enfoque postula un solo conjunto básico de estados de información que unifica aquellos dos. Podríamos decir que los aspectos internos de estos estados son fenoménicos, y que los aspectos externos son físicos. O a la manera de un eslogan: la experiencia es información desde el interior; la física es información desde el exterior.

## ¿Qué hay de la fenomenología macroscópica?

Todo esto funciona muy bien como ontología, aunque es ciertamente especulativo. Pero antes de dejarnos llevar demasiado, nos queda una enorme pregunta: ¿Cómo puede esta ontología hacerse compatible con los detalles de una *teoría* psicofísica? En particular, ¿cómo puede hacerse compatible con las regularidades psicofísicas en el nivel macroscópico? El problema es que el principio del doble aspecto se aplica aquí en el nivel físico fundamental, donde la información *microscópica* físicamente realizada tiene una realización fenoménica. Pero, para los propósitos de una teoría de la conciencia, necesitamos que la información *macroscópica* físicamente realizada tenga también una realización fenoménica. No es obvio en absoluto que este tipo de "fenomenología macroscópica" pueda derivarse de la fenomenología microscópica.

A primera vista, nuestra experiencia consciente no parece ser ningún tipo de suma de propiedades microfenoménicas correspondientes, por ejemplo, a las características físicas fundamentales en nuestro cerebro. Nuestra experiencia parece mucho más holística que eso y mucho más homogénea de lo que cualquier simple suma lo sería. Esto es una versión del "problema de la granularidad", planteado por Sellars (1965) como un problema del materialismo: ¿Cómo puede una experiencia ser idéntica a una vasta colección de sucesos fisiológicos, dada la homogeneidad de la primera y la fina granularidad de los segundos? El problema análogo es particularmente acuciante para los enfoques russellianos del tipo que estoy analizando.[6] Si las raíces de la fenomenología se agotan en la microfenomenología, entonces es difícil ver cómo podría surgir una fenomenología macroscópica estructurada y homogénea: lo que podríamos esperar, en cambio, es algún tipo de colección fenoménica no estructurada y "heterogénea".

Existen varios modos en los que podríamos intentar resolver esta cuestión. Primero, podríamos tratar de formular las cosas de modo que la ontología de doble aspecto sea válida en todos los niveles, no exclusivamente en el nivel microscópico. Es decir, aun los espacios de información físicos en el nivel macroscópico se basan en una realización fenoménica. Puede argumentarse que no hay nada privilegiado en el nivel microscópico: las cosas son más simples allí, pero no tienen por qué ser ontológicamente especiales. Los argumentos que hemos dado para ver el mundo físico en términos de información también se aplican en el nivel macroscópico. Podríamos argumentar incluso que en este nivel, sólo hay un espacio de diferencias macroscópicas que hacen una diferencia, cada una de las cuales podría realizarse en la fenomenología correspondiente.

El problema es que podría no haber espacio para todas estas realizaciones fenoménicas separadas. Una vez que tenemos características físicas fundamentales realizadas en espacios de información fenoménicos, entonces la información macroscópica parece ya tener una base: las diferencias que hacen una diferencia aquí están ahora basadas en configuraciones de características físicas microscópicas, que están a su vez basadas en la microfenomenología. Podríamos intentar introducir un basamento fenoménico separado de todas formas, pero esto parecería redundante, y teóricamente menos elegante que la maniobra correspondiente en el caso microscópico. Podríamos intentar remover la redundancia haciendo que el basamento macroscópico sea *primordial*, pero entonces sería difícil manejar casos de sistemas microscópicos aislados y similares. De modo que no es claro que el enfoque de "basamento" de la ontología del doble aspecto pueda funcionar directamente en el nivel macroscópico.

Segundo, podríamos intentar comprender algún modo en el cual la fenomenología macroscópica pudiese estar *constituida* por estas propiedades microfenoménicas. A primera vista, no parece ser ninguna simple suma o colección de estas propiedades; esto llevaría directamente a problemas de "irregularidades". Pero, tal vez, el problema consiste solamente en que no comprendemos la relación mental parte-todo, como Nagel (1986) lo formuló. Es decir, carecemos de una concepción precisa de la manera en el cual las propiedades microfenoménicas de bajo nivel "se componen" para producir una fenomenología de alto nivel. Tendemos a pensar acerca de la cuestión en términos de una analogía física basada en el modo como la microfísica compone la macrofísica, pero este podría ser un modo erróneo de pensar en ello. Quizá la fenomenología se constituya de un modo totalmente diferente.

Por ejemplo, podría ocurrir que las propiedades microfenoménicas compongan la macrofenomenología de una forma que refleja su estructura *informacional* conjunta, en lugar de su estructura espaciotemporal conjunta. Si una colección de estas propiedades realiza conjuntamente un estado complejo de información en virtud de las relaciones causales entre ellas, quizá podríamos esperar que cualquier macrofenomenología derivada tenga la forma de ese estado de información. Después de todo, el papel central de las propiedades microfenoménicas es realizar estados de información, de modo que no sería enteramente sorprendente que la estructura informacional desempeñe un papel en las relaciones constitutivas entre las propiedades. Si esto fuera así, entonces cualquier estado macrofenoménico derivado tendría la estructura informacional "homogénea" que el principio original del doble aspecto predice. Esto no es *fácil* de

comprender, pero después de todo no podemos esperar que nuestra comprensión cotidiana del dominio físico se aplique al dominio fenoménico. De modo que podría ocurrir que una mejor comprensión de la naturaleza de la propia fenomenología sea compatible con este enfoque de su constitución.

Si resultase que ninguna relación de constitución puede funcionar de este modo, podríamos intentar la tercera opción, que es vincular la macrofenomenología con la microfenomenología mediante *leyes*.[7] Por ejemplo, podría simplemente ser una ley que cuando los estados microfenoménicos realizan un estado de información de un cierto tipo en virtud de las relaciones causales entre ellos (por el principio de la "diferencia que hace una diferencia"), entonces surgirá una realización fenoménica directa del mismo estado. Esto resolvería los problemas teóricos, al costo de complicar la ontología. Ya no tendríamos una ontología simple en la que la fenomenología sería el aspecto intrínseco de la información físicamente realizada: alguna fenomenología quedaría "colgando" de esa información debido a las leyes, al modo de un dualismo de propiedades más estándar. De esta forma, algo del atractivo de un punto de vista russelliano se perdería, aunque el enfoque todavía sería bastante coherente.

Sea como fuere, dejaré esta cuestión abierta. Es el problema más difícil para cualquier clase de enfoque russelliano; pero no es obvio que no pueda ser resuelto. Si se pudiese hacer funcionar, la segunda estrategia parece una vía particularmente prometedora; también podría ocurrir que se necesite alguna idea enteramente nueva para resolver este problema. Si se miran las cosas de un modo optimista, podemos considerar el problema —cómo hacer que una teoría psicofísica sea compatible tanto con los hechos de macronivel acerca de nuestra fenomenología y su base física como con la ontología de micronivel de la perspectiva russelliana— como una de las restricciones cruciales que eventualmente podría llevarnos a una teoría detallada de la conciencia. Una de las dificultades para construir una teoría de este tipo es que no tenemos muchas restricciones. Podría ocurrir que este problema nos proporcione un muy necesario centro de atención.

Si ninguna de estas estrategias resulta ser satisfactoria, deberemos deshacernos del enfoque russelliano y utilizar alguna otra perspectiva de la metafísica. Podríamos intentar trabajar con la metafísica de la pura información, por ejemplo, como una manera de comprender el mundo físico; y luego de alguna manera conectar la fenomenología, tal vez mediante una conexión legaliforme, con la pura información. O podríamos simplemente retroceder a una ontología "domesticada" con dominios físico y fenoménico separados, cada uno con su propia naturaleza intrínseca, vinculados mediante co-

nexiones legaliformes del estilo del principio de información. Esto significaría que el discurso del "doble aspecto" debería tomarse menos seriamente y la ontología sería algo menos elegante, pero todavía podría conducirnos a una teoría perfectamente satisfactoria.

## 6. Preguntas abiertas

El esquema que presenté de un marco conceptual informacional de las leyes psicofísicas deja abiertas una enorme cantidad de preguntas. Para que la imagen pueda volverse una teoría final, todas estas preguntas deberán ser respondidas. En el apartado anterior mencioné algunos problemas con la ontología del enfoque. Pero también existen numerosas preguntas acerca de la forma de las *leyes* y acerca de cómo debe explicarse nuestra fenomenología. Algunas de estas preguntas son:

1. Cuando un espacio de información se realiza fenoménicamente, ¿por qué se realiza de un modo y no de otro? Por ejemplo, dado que nuestro espacio de color fenoménico podría estar invertido, parece algo arbitrario que sea del modo que es. ¿Necesitamos agregar otras leyes o postular "constantes" contingentes para decidir la cuestión?

2. ¿El carácter de un espacio de información fenoménico está determinado por la estructura del espacio (o al menos determinado, salvo la posibilidad de inversiones)? Podría parecer, por ejemplo, que el espacio del color y el espacio del gusto son ambos espacios tri o tetradimensionales simples, pero tienen características muy diferentes a pesar de sus formas similares. Es posible que la similitud de las estructuras sea una ilusión y que cuando las sumergimos en una estructura más amplia —si consideramos las experiencias de color como parte de la estructura profunda y completa de las experiencias visuales, por ejemplo— la similitud desaparezca. Pero la pregunta se mantiene: ¿Es algo con las características aproximadas de nuestras experiencias de color el *único* modo como la información visual del color podría haberse proyectado en la fenomenología, o existe algún modo totalmente diferente? Sospecho que la respuesta podría estar más cerca de la primera alternativa que de la segunda, pero no es obvio en absoluto cómo podríamos argumentar en favor de esto.

3. Utilicé este marco conceptual principalmente para analizar las experiencias perceptuales simples, como las experiencias del color. No es obvio cómo podríamos extenderlo para tratar con experiencias más sutiles, como por ejemplo las experiencias emocionales complejas o la experiencia del pensamiento ocurrente. ¿Pueden hacerse estas extensiones?

4. ¿Qué tipo de estructura formal es la más apropiada para capturar la estructura de la información fenoménica? ¿Qué tipo de espacios topológicos son necesarios para capturar la estructura relacional de la experiencia? ¿Deberíamos movernos a un tipo más específico de estructura, tal como un espacio métrico o una variedad diferencial? La estructura combinatoria de una experiencia es aun más interesante: un continuo multidimensional simple es probablemente una gran simplificación de la estructura de un campo visual, por ejemplo. ¿Cómo podemos capturar mejor la estructura completa? ¿Debería la definición de un espacio de información modificarse con este propósito?

5. ¿Cómo, dentro de este enfoque, podemos explicar la *unidad* de la conciencia? Esto es, ¿qué hace que mis experiencias visuales, auditivas, etc., sean todas experiencias del mismo sujeto? Sospecho que la respuesta involucra el modo como se procesa la información pertinente, de modo que la unidad de la conciencia corresponde al hecho de que la información pertinente está disponible para ser integrada de cierto modo. Pero, cómo exactamente traducir esto no es claro.

6. ¿Cuáles, exactamente, son los criterios que determinan qué información en el cerebro corresponde a *mis* experiencias? ¿Existe algún camino causal particular, o algún tipo determinado de flujo causal, que sea relevante? Podemos suponer que algo como la disponibilidad directa para el control global desempeña aquí un papel central en la individualización de la información y de los caminos relevantes.

La existencia de todas estas preguntas muestra lo lejos que estas ideas esquemáticas están de ser una verdadera teoría. Otro modo de ver esto es notar lo lejos que estas ideas están de permitirnos *predecir* exactamente cuáles son las propiedades fenoménicas asociadas con un sistema físico a partir de las propiedades físicas del sistema. Tal como se la formuló, la idea carece de *fuerza* explicativa y predictiva importante: se la debe reforzar de un modo considerable para poder ser realmente útil.

Se necesitarían algunos nuevos aportes para transformar esta idea en una teoría satisfactoria. Tal vez pudiese ocurrir un avance importante a partir de la consideración del problema del apartado anterior: cómo relacionar la información fenoménica a escala macroscópica con el enfoque de la "propiedad intrínseca" de la información en la escala microscópica. Otro avance podría provenir de tratar de hallar una restricción que produzca la clase de espacios de información físicamente realizados que se realizan en la

experiencia. Otros podrían surgir de fuentes que no consideré en absoluto.

La idea podría resultar totalmente equivocada. Esto no me sorprendería; de hecho, creo que es más probable que la clave de una teoría fundamental se encuentre en otro lado. No obstante, formulé estas ideas porque necesitamos comenzar a pensar acerca de estas cuestiones y porque la consideración de aunque más no sea un ejemplo inadecuado en el género puede ser instructivo. También, espero que algunas de las ideas planteadas sobre la marcha —acerca de cómo explicar los juicios fenoménicos, acerca de la ubicuidad de la experiencia y acerca de la conexión entre la experiencia, la información y las propiedades intrínsecas de lo físico— puedan resultar útiles, aun cuando se traduzcan a un marco conceptual diferente. Tal vez una teoría más adecuada de la conciencia pueda compartir algunos aspectos generales de las ideas aquí formuladas, aun cuando los detalles puedan ser muy diferentes.

Suele decirse que el problema con las teorías de la conciencia de este tipo es que son demasiado especulativas e inverificables. Sin embargo, creo que el verdadero problema con la "teoría" que formulé es otro: es demasiado inespecífica en sus predicciones. *Si* tuviésemos una teoría de un nivel comparable de simplicidad que pudiese *predecir* todos los hechos específicos relacionados con nuestra experiencia —aunque más no sea los hechos familiares en el caso de primera persona— a partir de los hechos físicos relacionados con nuestro sistema de procesamiento, esto sería un logro notable y nos daría muy buenas razones para aceptarla como verdadera. En este momento no contamos con una teoría de esta clase, pero no hay ninguna razón para creer que sea imposible.

# PARTE IV

# APLICACIONES

# 9

# Inteligencia artificial fuerte

## 1. La conciencia de las máquinas

¿Podría una máquina ser consciente? ¿Podría un ordenador programado de un modo apropiado verdaderamente poseer una mente? Estas preguntas han sido el tema de una enorme cantidad de debates en las últimas décadas. El campo de la *inteligencia artificial* (o IA) se dedica en gran parte a la meta de reproducir la mentalidad en máquinas computacionales. Hasta ahora el progreso ha sido limitado, pero sus defensores argumentan que tenemos todas las razones para creer que eventualmente los ordenadores tendrán una mente. Al mismo tiempo, los oponentes argumentan que los ordenadores están limitados de un modo que los seres humanos no lo están, de modo que está fuera de cuestión que una mente consciente surja meramente en virtud de cómputos.

Las objeciones a la inteligencia artificial típicamente adoptan una de dos formas. Primero, tenemos las objeciones *externas*, que intentan establecer que los sistemas computacionales nunca podrían *comportarse* como sistemas cognitivos. Según estas objeciones, existen ciertas capacidades funcionales que los seres humanos tienen pero que ningún ordenador podrá nunca poseer. Por ejemplo, a veces se sostiene que como estos sistemas siguen reglas, no podrían exhibir la conducta creativa o flexible característica de los seres humanos (por ejemplo, Dreyfus, 1972). Otros argumentaron que los ordenadores nunca podrían duplicar la comprensión matemática humana, ya que los sistemas computacionales están limitados por el teorema de Gödel, de un modo que los seres humanos no lo están (Lucas, 1961; Penrose, 1989).

Las objeciones externas resultaron difíciles de sostener, dado el éxito de la simulación computacional de procesos físicos en general.

En particular, parecería que tenemos buenas razones para creer que las leyes de la física son computables, de modo que al menos deberíamos poder simular computacionalmente la conducta humana. A veces, esto se pone en duda, argumentando en favor de un elemento no computable en las leyes físicas (como lo hace Penrose) o argumentando en favor de una causalidad no física (como lo hace Lucas), pero es claro que los que formulan estas objeciones pelean una batalla cuesta arriba.

De mayor importancia han sido las que yo denomino objeciones *internas*. Estas objeciones conceden, al menos a los fines de la discusión, que los ordenadores podrían simular la conducta humana, pero argumentan que de cualquier forma no tendrían una vida interior: ninguna experiencia consciente, ninguna verdadera comprensión. Como máximo, un ordenador podría proporcionar una simulación de la mente, no una reproducción. La objeción más conocida de esta clase es el argumento del "cuarto chino" de John Searle (1980). Según estas objeciones, los sistemas computacionales tendrían cuanto más la cáscara hueca de una mente: serían versiones en silicio de un zombi.

Aquellos que adoptan un enfoque no reduccionista de la experiencia consciente suelen sentirse atraídos por las objeciones internas a la inteligencia artificial; muchos de ellos argumentan que ningún mero ordenador podría ser consciente. Algunos a quienes el problema de la conciencia impresionó caracterizaron el problema señalando a la conciencia como ¡la característica que nosotros tenemos pero que los ordenadores no poseen! Muchos encontraron difícil de creer que un sistema artificial no biológico pueda ser el tipo de cosa capaz de dar origen a la experiencia consciente.

Sin embargo, un enfoque no reduccionista de la conciencia no lleva automáticamente a un enfoque pesimista de la IA. Las dos cuestiones son bastante distintas. La primera concierne a la *fuerza* de la conexión entre los sistemas físicos y la conciencia: ¿La conciencia está constituida por procesos físicos o meramente surge a partir de procesos físicos? La segunda concierne a la *forma* de la conexión: ¿Exactamente *cuáles* sistemas físicos dan origen a la conciencia? No es *obvio* que la ejecución del tipo correcto de cómputo deba dar origen a la conciencia; pero tampoco es obvio que los procesos neuronales en un cerebro deban hacerlo. A primera vista, no hay ninguna razón clara de por qué los ordenadores deberían estar en peores condiciones que los cerebros en este aspecto. Dado que hemos aceptado el hecho sorprendente de que los cerebros dan origen a la conciencia, no sería una *nueva* clase de sorpresa encontrar que el cómputo podría también




*Figura 9.1.* Bloom County sobre la IA fuerte. (© 1985, Washington Post Writers Group. Reproducido con autorización.)

darle origen. De modo que la mera aceptación de un enfoque no reduccionista de la conciencia debería dejar abierta la cuestión.

En este capítulo, llevaré las cosas aun más lejos y argumentaré que la ambiciones de la inteligencia artificial son razonables (fig. 9.1). En particular, argumentaré en favor de lo que Searle llama *inteligencia artificial fuerte*: que existe una clase no vacía de cómputos tal que la implementación de cualquier computación en esa clase es suficiente para la existencia de una mente y, en particular, es suficiente para la existencia de la experiencia consciente. Por supuesto, esta suficiencia sólo es válida con necesidad natural: es *lógicamente* posible que cualquier computación pueda ocurrir con ausencia de conciencia. Pero, como hemos visto, lo mismo ocurre para el cerebro. Al evaluar la perspectiva de la conciencia de las máquinas en el mundo real, es la posibilidad y necesidad natural lo que nos interesa.

(A riesgo de que esta conclusión sea considerada una trivialidad, dadas las sugerencias panpsíquicas del último capítulo, hago notar que nada en este capítulo se basa en esas consideraciones. No sólo argumentaré que es suficiente implementar la computación correcta para obtener conciencia, sino que la implementación de una computación correcta es suficiente para obtener una experiencia consciente rica como la nuestra.)

Ya realicé la mayor parte del trabajo necesario para esta defensa de la IA fuerte cuando argumenté, en el capítulo 7, en favor del principio de invariancia organizacional. Si ese argumento es correcto, establece que cualquier sistema con el tipo correcto de organización funcional es consciente, independientemente de la sustancia de la que esté hecho. De esta forma, ya sabemos que estar hecho de silicio, por ejemplo, no es ningún impedimento para poseer conciencia. Lo que nos queda por hacer es aclarar el vínculo entre la computación y la

organización funcional para establecer que la implementación de una computación apropiada es suficiente para asegurar la presencia de la organización funcional pertinente. Hecho esto, la IA fuerte surge como consecuencia. También responderé algunas objeciones que se han planteado en contra de la empresa de la IA fuerte.

## 2. La implementación de una computación

En su forma estándar, la teoría de la computación se ocupa totalmente de objetos *abstractos*: máquinas de Turing, programas en Pascal, autómatas de estados finitos, etc. Estas son entidades matemáticas que habitan el espacio matemático. Los sistemas cognitivos en el mundo real, por otro lado, son objetos *concretos*, corporizados físicamente y que interactúan causalmente con otros objetos en el mundo físico. Pero frecuentemente queremos utilizar la teoría de la computación para extraer conclusiones acerca de objetos concretos en el mundo real. Para hacer esto, necesitamos un puente entre los dominios abstracto y concreto.[1]

Este puente es la noción de *implementación*: la relación entre objetos computacionales abstractos —"computaciones" para abreviar— y sistemas físicos que ocurre cuando un sistema físico "realiza" una computación, o cuando una computación "describe" un sistema físico. Las computaciones suelen implementarse en ordenadores sintéticos basados en silicio, pero también pueden implementarse de otros modos. Por ejemplo, se dice frecuentemente que los sistemas naturales como el cerebro humano implementan computaciones. Las descripciones computacionales se utilizan para comprender los sistemas físicos en todo tipo de dominios. Cada vez que esto ocurre, es una noción de implementación la que, en forma implícita o explícita, hace el trabajo.

Sin embargo, rara vez se analiza en detalle la noción de implementación; en general simplemente se la presupone. Pero, para defender una IA fuerte, necesitamos una concepción detallada de ella. La tesis de la IA fuerte se formula en términos de computación, y nos dice que la implementación de la computación apropiada es suficiente para la conciencia. Para evaluar esta aseveración, debemos saber exactamente qué significa que un sistema físico implemente una computación. Una vez que sabemos esto, podemos combinarlo con nuestros análisis anteriores de las leyes psicofísicas para determinar si podría deducirse la conclusión.

Algunos argumentaron que no es posible formular ninguna concepción útil de implementación. En particular, Searle (1990b) sostuvo que la implementación no es una cuestión objetiva sino, más

bien, "relativa al observador": puede considerarse que cualquier sistema implementa cualquier computación si se lo interpreta del modo apropiado. Searle afirma, por ejemplo, que es posible considerar que una pared implementa el programa de procesamiento de texto Wordstar. Si Searle tuviera razón, sería difícil ver cómo las nociones computacionales podrían tener algún papel fundamental en una teoría que, en última instancia, trata de sistemas concretos. En lo que respecta a la IA, estaría vacía de contenido o implicaría una forma fuerte de panpsiquismo. Pero creo que este tipo de pesimismo está equivocado: puede formularse fácilmente una concepción objetiva de implementación. En este apartado esbozaré una concepción de este tipo. (La exposición es un poco técnica, pero el resto del capítulo debería tener sentido aun si se omiten los detalles.)

Cualquier concepción de qué significa que una computación sea implementada dependerá de la clase de computaciones en cuestión. Existen muchos formalismos computacionales diferentes, con distintas clases de computaciones en correspondencia: máquinas de Turing, autómatas de estados finitos, programas en Pascal, redes conexionistas, autómatas celulares, etc. En principio, necesitamos una concepción de implementación para cada uno de estos formalismos. Aquí expondré una concepción de implementación para un solo formalismo, el de los *autómatas de estados combinatorios*. Esta clase de computaciones es suficientemente general, de modo que la concepción asociada de implementación puede extenderse fácilmente para aplicarse a otras clases.

Un autómata de estados combinatorios es un primo más sofisticado de un autómata de *estados finitos*. Un autómata de estados finitos (FSA: finite state automaton) se define mediante un conjunto finito de entradas, un conjunto finito de estados internos, un conjunto finito de salidas y un conjunto asociado de relaciones de *transición de estados*. Un estado interno de un FSA es un elemento simple $S_i$ que no posee ninguna estructura interna; lo mismo ocurre para las entradas y las salidas. Las relaciones de transición de estados especifican, para todo par posible de entrada y estado interno, un nuevo estado interno y una salida. Si el estado inicial de un FSA está definido, estas relaciones de transición de estados especifican cómo evolucionará en el tiempo y qué salidas producirá, dependiendo de qué entradas reciba. La estructura computacional de un FSA consiste de este conjunto relativamente simple de relaciones de transición de estados entre un conjunto de estados no estructurados.

Los autómatas de estados finitos son inadecuados para representar la estructura de la mayoría de las computaciones que son relevantes en la práctica, ya que los estados y las relaciones de

transición de estados en estas computaciones tienen, por lo general, una estructura interna compleja. Ninguna descripción FSA puede capturar toda la estructura presente en un programa en Pascal, por ejemplo, o en una máquina de Turing, o en un autómata celular. Es, por lo tanto, más útil concentrarse en una clase de autómatas que tengan estados internos estructurados.

Los autómatas de estados combinatorios (CSA: combinatorial-state automata) son como los FSA, excepto que sus estados internos están estructurados. Un estado de un CSA es un *vector* $[S^1, S^2, ..., S^n]$. Este vector puede ser finito o infinito, pero me concentraré en el caso finito. Los elementos de este vector pueden pensarse como los *componentes* del estado interno; corresponden a las células en un autómata celular o a las celdas de cinta y estado de la cabeza lectora en una máquina de Turing. Cada elemento Si, puede adoptar un número finito de valores $S^i_j$, donde $S^i_j$ es el valor posible j-ésimo del elemento i-ésimo. Pueden pensarse estos valores como "subestados" del estado general. Las entradas y salidas tienen un tipo similar de estructura compleja: una entrada es un vector $[I^1, ..., I^k]$ y una salida un vector $[O^1, ..., O^m]$.

Un CSA se define especificando el conjunto de vectores de estados internos, los vectores de entrada y salida, y un conjunto de *reglas de transición de estados* que determinan cómo el estado del CSA evoluciona en el tiempo. Para cada elemento del vector del estado interno, una regla de transición de estados determina cómo su nuevo valor depende de valores anteriores de los vectores de entrada y estado interno. Para cada elemento del vector de salida, una regla de transición de estados determina cómo su nuevo valor depende de los anteriores valores del vector de estado interno. Todo CSA finito puede representarse como un FSA con igual potencia computacional, pero la descripción del FSA sacrificará la mayor parte de la estructura que es crucial para un CSA. Esa estructura es fundamental para utilizar los CSA para capturar la organización que subyace a una mente.

Estamos ahora en posición de ofrecer una concepción de la noción de implementación. Las computaciones como los CSA son objetos abstractos, con una *estructura formal* determinada por sus estados y relaciones de transición de estados. Los sistemas físicos son objetos concretos, con una *estructura causal* determinada por sus estados internos y las relaciones causales entre los estados. Informalmente, decimos que un sistema físico *implementa* una computación cuando la estructura causal del sistema refleja la estructura formal de la computación. Es decir, el sistema implementa una computación si existe un modo de poner en correspondencia los estados del sistema

con los estados de la computación, de manera que los estados físicos que están causalmente relacionados se apliquen en estados formales que están formalmente relacionados del modo correspondiente.

Esta idea intuitiva puede aplicarse directamente para producir una concepción de la implementación de los CSA. Un sistema físico implementa un CSA si existe una descomposición de los estados internos del sistema en subestados, una descomposición de las entradas y salidas del sistema en subestados de entrada y de salida, y una aplicación de los subestados del sistema en subestados del CSA, tal que las relaciones causales de transición de estados entre estados físicos, entradas y salidas reflejan las relaciones formales de transición de estados entre los estados formales, las entradas y las salidas correspondientes.

El criterio formal para implementar un CSA es como sigue:

> Un sistema físico *P* implementa un CSA *M* si existe una descomposición de estados internos de *P* en componentes $[s^1,...,s^n]$ y una aplicación f desde los subestados sj en los subestados correspondientes *S*j de M, junto con descomposiciones y aplicaciones similares para las entradas y salidas, tal que para cada regla de transición de estados $([I^1,..., I^k], [S^1,..., S^n]) \rightarrow ([S'^1, ..., S'^n], [O^1,..., O^l])$ de *M*: si *P* es un estado interno $[s^1,..., s^n]$ y recibe la entrada $[i^1,..., i^n]$, que aplica en el estado formal y de entrada $[S^1,..., S^n]$ y $[I^1,..., I^k]$ respectivamente, esto causa fiablemente que pase a un estado interno y produzca una salida que se aplique en $[S'^1,..., S'^n]$ y $[O^1,..., O^l]$ respectivamente.

Podemos estipular que en una descomposición del estado de un sistema físico en un vector de subestados, el valor de cada elemento del vector debe supervenir en una región separada del sistema físico, para asegurar que la organización causal relacione componentes distintos del sistema. De otra manera, no es claro que la estructura causal detallada esté realmente presente dentro del sistema físico. Hay espacio para jugar con esto y con otros detalles en la definición de más arriba. La noción de implementación no está escrita en la piedra; podría restringirse o relajarse para diversos propósitos. Pero esto produce la forma básica que será compartida por cualquier concepción de implementación.

Podría parecer que los CSA no son un gran avance respecto de los FSA. Después de todo, para cualquier CSA finito, podemos encontrar un FSA correspondiente con la misma conducta de entrada-salida. Pero existen algunas diferencias cruciales. Primero y principal, las condiciones de *implementación* sobre un CSA son mucho más restrin-

gidas que las del FSA correspondiente. Una implementación de un CSA debe consistir de una interacción causal compleja entre un número de partes separadas; una descripción mediante un CSA puede entonces capturar la organización causal de un sistema en un grano mucho más fino. Segundo, los CSA proporcionan una concepción unificada de las condiciones de implementación para máquinas finitas e infinitas. Tercero, un CSA puede directamente reflejar la organización formal compleja de objetos computacionales como máquinas de Turing y autómatas celulares. En el FSA correspondiente, gran parte de esta estructura se perdería.

Podemos utilizar esta definición de implementación para directamente proporcionar criterios de implementación para otros tipos de computaciones. Para especificar lo que se necesita para implementar una máquina de Turing, por ejemplo, sólo necesitamos redescribir esta última como un CSA y aplicar la definición de más arriba. Para hacerlo, describimos el estado de la máquina de Turing como un vector gigante. Un elemento del vector representa el estado de la cabeza lectora de la máquina, y hay un elemento para cada celda de la cinta, que representa el símbolo en la celda y también indica si la cabeza lectora está en esa posición. Las reglas de transición de estados entre los vectores son las que se derivan naturalmente de los mecanismos que especifican la conducta de la cabeza lectora y la cinta. Por supuesto, los vectores aquí son infinitos, pero las condiciones de implementación en el caso infinito son una extensión directa de las del caso finito. Dada esta traducción del formalismo de la máquina de Turing al formalismo del CSA, podemos decir que se implementa una máquina de Turing cada vez que se implementa el CSA correspondiente. Podemos formular traducciones similares para las computaciones en otros formalismos, como los autómatas celulares o los programas en Pascal, lo que produce condiciones de implementación para las computaciones en cada una de esas clases.

Esto genera un criterio perfectamente objetivo para implementar una computación. La implementación de una computación no colapsa en el vacío como Searle sugiere. Es verdad que *algunas* computaciones serán implementadas por cualquier sistema. Por ejemplo, un CSA de un único estado y un único elemento puede ser implementado por cualquier sistema, y un CSA de dos estados podrá ser implementado casi tan ampliamente. También es verdad que la mayoría de los sistemas implementarán más de una computación, según cómo dividamos los estados del sistema. No hay nada sorprendente en esto: es de esperar que mi estación de trabajo implemente un número de computaciones como lo hace mi cerebro.

Lo crucial es que no hay razones para creer que *todo* CSA pueda ser implementado por *todo* sistema. Para cualquier CSA complejo dado, muy pocos sistemas físicos tendrán la organización causal necesaria para implementarlo. Si tomamos un CSA cuyos vectores de estado tienen mil elementos, con diez posibilidades para cada uno, entonces argumentos similares a los presentados en el capítulo 7 sugieren que la probabilidad de que un conjunto arbitrario de estados físicos tenga las relaciones causales requeridas es algo menos de uno en $(10^{1000})^{10^{1000}}$ (en realidad mucho menos que esto, debido al requerimiento de que las relaciones de transición sean fiables).[2]

¿Qué hay de la afirmación de Searle de que las descripciones computacionales son "relativas al observador", entonces? Es verdad que existe un grado limitado de relatividad del observador: cualquier sistema físico particular implementará un número de computaciones, y en cuál de ellas decida concentrarse un observador dependerá de sus propósitos. Pero esto no es una amenaza para la IA o la ciencia cognitiva computacional. Sigue siendo el caso que para cualquier computación dada, es una cuestión fáctica si un sistema particular la implementa o no, y sólo habrá una clase limitada de sistemas que califiquen como implementaciones. Para que las concepciones computacionales puedan tener injerencia metafísica y explicativa, eso es todo lo que esas disciplinas necesitan.

Decir que un sistema físico implementa una computación compleja determinada *P* es decir algo muy sustancial acerca de su estructura causal, algo que podría ser bastante útil para proporcionar explicaciones cognitivas y, tal vez, para comprender la base de la conciencia. Sólo sistemas con un tipo muy específico de organización causal podrán tener la esperanza de satisfacer las fuertes restricciones de la implementación. De modo que no hay ningún peligro de vacuidad y, en cambio, podemos esperar que la noción de computación proporcione una base sustancial para el análisis de los sistemas cognitivos.

## 3. En defensa de una IA fuerte

Lo que se requiere para implementar un CSA es notablemente similar a lo que se requiere para realizar una organización funcional. Recuérdese que una organización funcional se determina especificando un número de componentes abstractos, un número de estados para cada componente y un sistema de relaciones de dependencia que indican cómo los estados de cada componente dependen de los estados previos y de las entradas, y cómo las salidas dependen de los estados previos. La noción de un CSA es efectivamente una formalización directa de esta noción.

Dada cualquier organización funcional del tipo descripto en el capítulo 7, podemos fácilmente abstraerla en un CSA. Sólo necesitamos estipular que los vectores de estados del CSA tengan un elemento para cada componente de la organización y que las transiciones formales de estados entre los estados del CSA correspondan a las relaciones causales de dependencia entre los componentes. Realizar la organización funcional se reduce casi exactamente a lo mismo que implementar el CSA correspondiente. Existen algunas pequeñas diferencias, tales como diferentes tratamientos de las entradas y salidas, pero estas no son significativas.

La concepción de la implementación que formulé aclara entonces el vínculo entre la organización causal y la computacional. De este modo, podemos ver que cuando las descripciones computacionales se aplican a sistemas físicos, proporcionan efectivamente una descripción formal de la organización causal del sistema. El lenguaje de la computación nos proporciona un lenguaje perfecto en el cual puede especificarse este tipo de organización causal abstracta. Es posible argumentar que esta es precisamente la razón por la cual las nociones computacionales tienen una aplicación tan amplia en toda la ciencia cognitiva. Lo más importante para la explicación de la conducta de un sistema cognitivo complejo es la organización causal abstracta del sistema, y los formalismos computacionales proporcionan un marco ideal dentro del cual puede describirse y analizarse este tipo de organización.[3]

Este vínculo hace que la defensa de una inteligencia artificial fuerte sea simple. Ya argumenté en favor del principio de la invariancia organizacional, que nos dice que para cualquier sistema con experiencias conscientes, un sistema con la misma organización funcional detallada tendrá experiencias conscientes cualitativamente idénticas. Pero sabemos que cualquier organización funcional determinada puede abstraerse en un CSA que se implementa cada vez que se realiza la organización. Se deduce que para un sistema consciente particular $M$, su organización funcional de grano fino puede abstraerse en un CSA $M$, tal que cualquier sistema que implemente $M$ realizará la misma organización funcional y, por lo tanto, tendrá experiencias cualitativamente indistinguibles de las del sistema original. Esto establece la tesis de una inteligencia artificial fuerte.

Por ejemplo, podríamos abstraer una descripción neuronal del cerebro en un CSA, con un elemento del vector de estados por cada neurona y con subestados para cada elemento que reflejan el conjunto pertinente de estados de cada neurona. Las reglas de transición de estados del CSA reflejan el modo como el estado de cada neurona depende del estado de otras neuronas y el modo como los estados

neuronales están relacionados con entradas y salidas. Si los componentes no neuronales del cerebro son relevantes, podemos también incluir elementos para ellos. Cualquier sistema físico que implemente este CSA tendrá una organización funcional de grano fino que duplica la organización funcional de nivel neuronal del cerebro. Por el principio de invariancia, este sistema tendrá experiencias indistinguibles de las asociadas con el cerebro.

Es fácil pensar en un ordenador como simplemente un dispositivo de entrada-salida con nada en medio excepto algunas manipulaciones matemáticas formales. Este enfoque, sin embargo, deja afuera el hecho clave de que existe una dinámica causal rica dentro de un ordenador, así como existe en el cerebro. En un ordenador ordinario que implemente una simulación de mi cerebro neurona por neurona, habrá una causalidad real entre voltajes en diversos circuitos que reflejará precisamente los patrones de causalidad entre las neuronas. Por cada neurona, habrá una localización de memoria que representa a esa neurona y cada una de estas localizaciones se realizará físicamente en un voltaje en alguna localización física. Son los patrones causales entre estos circuitos, así como los son los patrones causales entre las neuronas en el cerebro, los responsables de cualquier experiencia consciente que surja.

También podemos defender la tesis de la IA fuerte directamente, utilizando los argumentos de los qualia gradualmente desvanecientes y danzantes. Dadas dos implementaciones de un CSA existirá un espectro de casos entre ellos, en los cuales los componentes físicos de las implementaciones se reemplazan uno por vez mientras que el patrón de su interacción causal con el resto del sistema se preserva. Si uno de los sistemas es consciente, y si el CSA abstrae su organización funcional de grano fino, entonces los argumentos en cuestión implican que el otro sistema debe ser consciente y que debe tener experiencias conscientes indistinguibles. Si el otro sistema no fuese consciente, habría un estado intermedio con qualia desvanecientes. Si el otro sistema no fuese consciente o tuviese experiencias conscientes diferentes, entonces podríamos construir un sistema intermedio con qualia danzantes. Estas consecuencias no son plausibles, por las razones expuestas en el capítulo 7. Dado que los qualia no pueden desvanecerse o danzar de ese modo, se deduce que el segundo de los sistemas originales tiene experiencias indistinguibles del primero, y que la tesis de la IA fuerte se mantiene.

Existe una pequeña salvedad. El argumento supone que la organización del cerebro puede abstraerse en la descripción de un CSA. Esto requiere solamente que pueda describirse la organización relevante en términos de un número finito de partes, cada una con un

número finito de estados pertinentes. Sin embargo, algunos podrían no estar de acuerdo con esto. Por ejemplo, tal vez sea necesario un número infinito de estados para cada neurona, para capturar el papel vital del procesamiento continuo. Algunos podrían afirmar que las transiciones entre estos estados infinitos podrían ser no computables. Analizaré este tipo de objeción más adelante; por ahora, me contentaré con aceptar la conclusión de que *si* la dinámica cognitiva es computable, entonces el tipo correcto de organización computacional dará origen a la conciencia. Esto es, estoy más interesado por las objeciones internas que por las externas. Aun así, analizaré algunas objeciones externas más adelante en el capítulo.

## 4. El cuarto chino y otras objeciones

Por supuesto, los opositores de la IA fuerte a veces formularon argumentos concretos en contra de esta posición. El mejor conocido de estos se debe a John Searle, en su artículo de 1980, "Minds, Brains, and Programs", y en otros lados. Aquí utilizaré el marco conceptual que desarrollé antes para hacer frente a estos argumentos.

### El cuarto chino

En un celebrado argumento en contra de la IA fuerte, Searle (1980) sostiene que cualquier programa puede implementarse sin que ello dé origen a una mente. Hace esto exhibiendo lo que interpreta que es un *contraejemplo* de la aseveración de la IA fuerte: el cuarto chino, dentro del cual una persona que manipula símbolos simula a alguien que comprende el chino. El sentido del cuarto chino es proporcionar un ejemplo, para cualquier programa particular, de un sistema que implementa ese programa pero que carece de la experiencia consciente pertinente.

En la versión original, Searle dirige el argumento en contra de la *intencionalidad* de las máquinas, en lugar de en contra de la conciencia de las mismas, argumentando que es "comprensión" lo que el cuarto chino no posee. A pesar de todo, es bastante claro que lo que está en la raíz del problema es la conciencia. De ser exitoso, lo que el núcleo de su argumento establecería directamente es que el sistema del cuarto chino carece de estados conscientes, como la experiencia consciente de comprender el chino. Según el enfoque de Searle, la intencionalidad requiere la conciencia, de modo que esta es suficiente para ver que el cuarto carece también de intencionalidad. Otros lo niegan, sin embargo. De cualquier forma, podemos hacer abstracción

del tema de la conexión entre la conciencia y la intencionalidad, y formular la cuestión sólo en términos de la primera. De esta manera, los problemas podrían ser más claros.

(Es decir, podemos separar las conclusiones de Searle en dos partes: 1) ningún programa es suficiente para la conciencia; y 2) ningún programa es suficiente para la intencionalidad. Searle cree que 1) implica 2), pero otros lo niegan. Las cuestiones resultan más claras si se interpreta que el argumento acerca de la IA fuerte se concentra en 1): todas las partes aceptarán que si 1) es verdad, entonces la forma más interesante de la IA fuerte está condenada, e incluso Searle aceptaría que la refutación de 1) mostraría que el argumento del cuarto chino es fallido. El vínculo entre la conciencia y la intencionalidad puede entonces dejarse de lado como una cuestión separada, no crucial para el argumento en contra de la IA.

De esta manera, evitamos la situación en la que los opositores argumentan en contra de 2) sin preocuparse por argumentar en contra de 1). Por ejemplo, las respuestas que se concentran en la conexión entre el cuarto chino y su entorno [Fodor, 1980; Rey, 1986] y las respuestas que ofrecen concepciones procesales o funcionales de la intencionalidad [Boden, 1988; Thagard, 1986] pueden o no arrojar luz sobre la cuestión de la intencionalidad, pero no contribuyen en nada a la plausibilidad de la conciencia del cuarto chino. Por consiguiente, nos dejan con la sensación de que no se ocuparon del problema que este escenario plantea para la IA. En el mejor de los casos, lo que estuvo en discusión fue la premisa auxiliar de que la intencionalidad requiere la conciencia.)[4]

El argumento del cuarto chino se desarrolla de la siguiente manera. Tómese cualquier programa que se supone captura algún aspecto de la conciencia, como la comprensión del chino o tener una sensación de rojo. Este programa puede ser implementado por un hablante monolingüe del inglés —aquí lo llamaremos el *demonio*— en una habitación en blanco y negro. El demonio sigue manualmente todas las reglas especificadas por el programa, mantiene un registro de todos los estados internos y las variables pertinentes en tiras de papel, y las borra y actualiza según las necesidades. Podemos imaginar que el demonio está conectado también con un cuerpo robot que recibe las entradas digitales de los transductores perceptuales, los manipula de acuerdo con las especificaciones del programa y envía salidas digitales a efectores motores. De este modo, el programa se implementa perfectamente. Sin embargo, es evidente que el demonio no comprende conscientemente el chino y que tampoco experimenta una sensación de rojo. Por lo tanto, implementar un programa no es

suficiente para la existencia de estas experiencias conscientes. La conciencia debe requerir algo más que la implementación de un programa apropiado.

Típicamente, los defensores de la IA fuerte respondieron aceptando que el *demonio* no comprende el chino y argumentando que la comprensión y la conciencia deberían atribuirse en cambio al *sistema* consistente del demonio y los trozos de papel. Searle declaró que esta réplica es manifiestamente poco razonable. Es cierto que hay algo poco intuitivo en la afirmación de que un sistema compuesto por un agente y trozos de papel asociados tiene una conciencia colectiva. En este punto, el argumento alcanza un callejón sin salida. Los defensores de la IA sostienen que el sistema es consciente, los opositores encuentran que la conclusión es ridícula, y parece difícil ir más allá. Creo que los argumentos ya enunciados nos dan una base para romper el impasse en favor de una IA fuerte, sin embargo.

Supongamos que el programa pertinente es, de hecho, un autómata de estados combinatorios que refleja la organización de nivel neuronal de un hablante chino que mira una jugosa manzana roja. El demonio en el cuarto implementa el CSA: mantiene una tira de papel para cada elemento del vector de estados y actualiza las tiras de papel en cada paso de acuerdo con las reglas de transición de estados. Podemos utilizar los argumentos de los qualia desvanecientes y danzantes y construir un espectro de casos entre el hablante chino original y el cuarto chino.[5] Esto no es difícil de hacer. Primero, podemos imaginar que las neuronas en la cabeza del hablante chino son reemplazadas una por vez por diminutos demonios, cada uno de los cuales reproduce la función de entrada-salida de una neurona.[6] Al recibir estimulación de las neuronas vecinas, un demonio hace los cálculos apropiados y estimula, a su vez, a las neuronas vecinas. A medida que se reemplazan cada vez más neuronas, los demonios asumen el control, hasta que el cráneo está lleno de miles de millones de demonios que reaccionan a las señales de los otros y a las entradas sensoriales, hacen cálculos y, a su vez, hacen señales a otros demonios y estimulan salidas motoras. (Si alguien objeta que todos esos demonios nunca podrían entrar en un cráneo, podemos imaginar un escenario con un equipo de transmisión de radio fuera del cráneo.)

Luego, gradualmente reducimos el número de demonios permitiéndoles duplicar su trabajo. Al principio reemplazamos dos demonios vecinos por un solo demonio que hace la tarea de los dos. El nuevo demonio mantendrá un registro del estado interno de las dos neuronas que simula; podemos imaginar que ese registro se conserva en un trozo de papel en cada localización. Cada trozo de papel será actualizado dependiendo de las señales desde los demonios vecinos y

también del estado del otro trozo de papel. Los demonios se consolidan aun más hasta que, eventualmente, sólo queda un demonio y miles de millones de pequeñas tiras de papel. Podemos imaginar que cada una de esas tiras está en la localización original de su neurona correspondiente y que el demonio corre alrededor del cerebro actualizando cada tira de papel en función de los estados de las tiras vecinas y de las entradas sensoriales, cuando es necesario.

A pesar de todos estos cambios, el sistema resultante comparte la organización funcional del cerebro original. Las relaciones causales entre neuronas en el caso original se reflejan en las relaciones causales entre los demonios en el caso intermedio, y en las relaciones causales entre las tiras de papel en el caso final. En este último, las relaciones causales son mediadas por las acciones de un demonio —un trozo de papel afecta el estado del demonio, quien a su vez afecta los trozos vecinos de papel— pero, de todas formas, son relaciones causales. Si observamos al sistema funcionar a una velocidad acelerada, veremos un ajetreo de interacción causal que corresponde precisamente al ajetreo entre las neuronas.

Podemos ahora aplicar los argumentos de los qualia desvanecientes y danzantes. Si el sistema final no posee experiencias conscientes, entonces debe haber un sistema intermedio con experiencias conscientes desvanecientes. Esto es inverosímil por las mismas razones que antes. Podemos también imaginar que conmutamos entre un circuito neuronal y un circuito de respaldo correspondiente implementado con demonios, o con un solo demonio y trozos de papel. Como antes, esto llevaría a qualia danzantes con una organización funcional constante, de modo que el sistema nunca podría notar la diferencia. Una vez más, es mucho más plausible suponer que los qualia permanecen constantes en todo momento.

Por lo tanto, es razonable concluir que el sistema final tiene precisamente las experiencias conscientes del sistema original. Si el sistema neuronal dio origen a experiencias de rojo brillante, también lo hará el sistema de demonios, y también la red de trozos de papel mediados por un demonio. Pero, por supuesto, este caso final es sólo una copia del sistema del cuarto chino. Por lo tanto hemos dado una razón positiva para creer que el sistema tiene experiencias conscientes, tal como las de comprender el chino o experimentar el rojo.

Este punto de vista deja en claro dos cosas que podrían resultar oscurecidas por la descripción de Searle del cuarto chino. Primero, las "tiras de papel" en la habitación no son una mera pila de símbolos formales. Constituyen un sistema dinámico concreto con una organización causal que corresponde directamente a la organización del cerebro original. El ritmo lento que asociamos con la manipulación de

símbolos oscurece esta cuestión, como también lo hace la presencia del demonio que manipula los símbolos pero, no obstante, es la dinámica concreta entre los trozos de papel lo que da origen a la experiencia consciente. Segundo, el papel del demonio es totalmente secundario. La dinámica causal interesante es la que ocurre entre los trozos de papel, que corresponden a las neuronas en el caso original. El demonio simplemente actúa como una especie de facilitador causal. La imagen de un demonio que corre a toda prisa alrededor del cráneo deja en claro que la atribución de las experiencias del sistema al *demonio* sería una seria confusión de niveles. El hecho de que el demonio es un agente consciente podría llevarnos a suponer que si las experiencias del sistema están en algún lado, están en el demonio; pero, de hecho, la conciencia del demonio es totalmente irrelevante para el funcionamiento del sistema. El trabajo del demonio podría ser realizado por una simple tabla de doble entrada. El aspecto crucial del sistema es la dinámica entre los símbolos.

El argumento de Searle se basa en nuestras intuiciones cuando implementamos el programa de un modo extravagante que oscurece la realización de la dinámica causal pertinente. Una vez que miramos más allá de las imágenes que nos produce la presencia del demonio irrelevante y la lenta velocidad de manipulación de símbolos, sin embargo, podemos ver que la dinámica causal en el cuarto refleja precisamente la dinámica causal en el cráneo. De este modo, ya no parece tan inverosímil suponer que el sistema pueda dar origen a la experiencia.

Searle también ofrece una versión del argumento en el cual el demonio *memoriza* las reglas de la computación e implementa el programa internamente. Por supuesto, en la práctica, las personas no pueden memorizar ni siquiera cien reglas y símbolos, y mucho menos muchos miles de millones, pero podemos imaginar que un demonio con un módulo de supermemoria podría ser capaz de memorizar todas las reglas y los estados de todos los símbolos. En este caso, podemos nuevamente esperar que el sistema de origen a experiencias conscientes que no son las experiencias del demonio. Searle sostiene que si alguien tiene las experiencias ese debe ser el demonio, ya que todo el procesamiento es interno al mismo; sin embargo, esto debería considerarse un ejemplo de dos sistemas mentales realizados dentro del mismo espacio físico. La organización que da origen a las experiencias del chino es bastante distinta de la organización que da origen a las experiencias del demonio. La organización de comprensión del chino radica en las relaciones causales entre miles de millones de localizaciones en el módulo de supermemoria; una vez más, el demonio sólo actúa como una especie de facilitador causal. Esto se hace

evidente si consideramos un espectro de casos en los cuales el demonio corre a toda velocidad a través del cráneo y gradualmente memoriza las reglas y símbolos hasta que todo está internalizado. La estructura pertinente se mueve gradualmente desde el cráneo a la supermemoria del demonio, pero la experiencia se mantiene constante en todo momento y separada completamente de sus experiencias.

Algunos podrían suponer que debido a que mi argumento se basa en duplicar la organización de nivel neuronal del cerebro, sólo establece entonces una forma débil de IA fuerte, una que está estrechamente ligada a la biología. (Al analizar lo que él llama la respuesta del "Simulador cerebral", Searle expresa sorpresa de que un defensor de la IA pudiera dar una respuesta que depende de la simulación detallada de la biología humana.) Sin embargo, esto significa no haber advertido cuál es la fuerza del argumento. El programa de simulación cerebral es meramente el extremo aguzado de la cuña. Una vez que sabemos que *un* programa puede dar origen a una mente aun cuando está implementado al estilo del cuarto chino, la fuerza del argumento principista de Searle resulta totalmente eliminada: sabemos que el demonio y los papeles en un cuarto chino pueden sustentar una mente independiente. Las esclusas se abren, entonces, a toda una variedad de programas que podrían ser candidatos para generar experiencias conscientes. La amplitud de esta variedad es una cuestión abierta, pero el cuarto chino ya no es un obstáculo.

## Sintaxis y semántica

Un segundo argumento formulado por Searle (1984) es el siguiente:

1. Un programa de ordenador es sintáctico.
2. La sintaxis no es suficiente para la semántica.
3. Las mentes tienen semántica.
4. Por lo tanto, implementar un programa es insuficiente para una mente.

Una vez más, el filósofo formula esto como un argumento acerca de la intencionalidad, pero también puede interpretarse como un argumento acerca de la conciencia. Para Searle, el tipo fundamental de intencionalidad es la intencionalidad fenomenológica, el tipo que es inherente a la conciencia.

Existen varios modos en los que puede interpretarse y criticarse este argumento, pero el problema principal es que no respeta el papel

crucial de la implementación. Los *programas* son objetos computacionales abstractos y son puramente sintácticos. Ciertamente, ningún mero programa es un candidato para la posesión de una mente. Las *implementaciones de programas*, por otro lado, son sistemas concretos con una dinámica causal y no son puramente sintácticos. Una implementación tiene peso causal en el mundo real y es en virtud de ese peso causal que surgen la conciencia y la intencionalidad. Es el programa el que es sintáctico; es la implementación la que tiene contenido semántico.

Searle podría argumentar que existe un sentido en el cual aun las implementaciones son sintácticas, tal vez porque la dinámica de las implementaciones está determinada por propiedades formales. Sin embargo, cualquier sentido de "sintaxis" en el cual las implementaciones son sintácticas, pierde contacto con el sentido en el cual es plausible que la sintaxis no sea suficiente para la semántica. Aunque puede ser plausible que conjuntos estáticos de símbolos abstractos no tengan propiedades semánticas intrínsecas, es mucho menos claro que los procesos causales formalmente especificados no puedan soportar una mente.

Podemos parodiar el argumento del siguiente modo:

1. Las recetas son sintácticas.
2. La sintaxis no es suficiente para la esponjosidad.
3. Las tortas son esponjosas.
4. Por lo tanto, implementar una receta es insuficiente para una torta.

En esta forma el defecto es inmediatamente evidente. El argumento no distingue entre recetas, que son objetos sintácticos, e implementaciones de recetas, que son sistemas físicos completos en el mundo real. Nuevamente, todo el trabajo lo hace la relación de implementación, que relaciona los dominios abstracto y concreto. Una receta especifica implícitamente una clase de sistemas físicos que califican como *implementaciones* de la receta; son estos sistemas los que tienen características como la de esponjosidad. De modo similar, un programa especifica implícitamente una clase de sistemas físicos que califican como implementaciones del programa y son estos sistemas los que dan origen a características como las mentes.

### Una simulación es sólo una simulación

Una objeción popular a la inteligencia artificial (por ejemplo, Searle, 1980; Harnad, 1989) es que una simulación de un fenómeno

no es lo mismo que una reproducción de él. Por ejemplo, cuando simulamos computacionalmente la digestión, no se digiere concretamente ninguna comida. Un huracán simulado no es un verdadero huracán; cuando se simula un huracán en un ordenador, nadie se moja. Cuando se simula el calor, no se produce ningún verdadero calor. De modo que cuando se simula una mente, ¿por qué deberíamos esperar que resulte una verdadera mente? ¿Por qué deberíamos esperar en este caso, pero no en los otros, que un proceso computacional no sea sólo una simulación sino la cosa real?

Es verdad que para muchas propiedades, una simulación no es una reproducción. El calor simulado no es verdadero calor. Por otro lado, para algunas propiedades, una simulación *es* una reproducción. Por ejemplo, una simulación de un sistema con un bucle causal *es* un sistema con un bucle causal. De modo que la verdadera pregunta es, ¿cómo distinguimos esos tipos *X* en los que una simulación de un *X* es realmente un *X* de aquellos en los que esto no ocurre?

Sugiero que la respuesta es la siguiente. Una simulación de *X* es un *X* precisamente cuando la propiedad de ser un *X* es un *invariante organizacional*. La definición de esta propiedad es como antes: una propiedad es un invariante organizacional cuando sólo depende de la organización funcional del sistema subyacente, y no de algún otro detalle. Una simulación computacional de un sistema físico puede capturar su organización causal abstracta y asegurarse de que esta organización se reproduzca en cualquier implementación, independientemente de en qué medio se la realice. Una implementación de esta clase *reproducirá* entonces cualquier invariante organizacional del sistema original, pero otras propiedades se perderán.

La propiedad de ser un huracán no es un invariante organizacional, ya que depende parcialmente de propiedades no organizacionales tales como la velocidad, la forma y la composición física del sistema subyacente (un sistema con las mismas interacciones causales implementadas muy lentamente entre un gran conjunto de bolas de billar no sería un huracán). De modo similar, la digestión y el calor dependen de aspectos de la organización física subyacente que no son del todo organizacionales. Podríamos sustituir gradualmente los componentes biológicos en un sistema digestivo de modo que las reacciones ácido-base se reemplacen por interacciones causalmente isomórficas entre trozos de metal, pero esto ya no sería una instancia de digestión. De modo que no deberíamos esperar que una simulación de sistemas con estas propiedades tenga ella misma esas propiedades.

Pero las propiedades fenoménicas son diferentes. Como argumenté en el capítulo 7, son invariantes organizacionales. Si esto es así, se deduce que el tipo correcto de simulación de un sistema con

propiedades fenoménicas tendrá él mismo propiedades fenoménicas, en virtud de reproducir la organización funcional de grano fino del sistema original. La invariancia organizacional hace que la conciencia sea diferente en principio de las otras propiedades mencionadas y abre el camino a la IA fuerte.

## 5. Objeciones externas

Hasta ahora me ocupé fundamentalmente de las objeciones internas a la inteligencia artificial fuerte, ya que ellas son las más relevantes al tema de este libro pero, al menos, mencionaré algunas objeciones externas. Como dije antes, el caso *prima facie* en contra de las objeciones externas a la inteligencia artificial es fuerte: tenemos todas las razones para creer que las leyes de la física, al menos tal como se las entiende actualmente, son computables, y que la conducta humana es una consecuencia de leyes físicas. Si esto es así, entonces se deduce que un sistema computacional puede simular la conducta humana. No obstante, de todos modos, ocasionalmente se plantean estas objeciones, de manera que las analizaré aquí en forma sintética.

### Objeciones a partir del seguimiento de reglas

Tal vez la objeción externa más antigua a la IA sea que los sistemas computacionales siempre siguen reglas, de modo que nunca poseerán ciertas capacidades humanas como la creatividad o la flexibilidad. Esta es, en muchos sentidos, la más débil de las objeciones externas, en parte debido a su vaguedad y subespecificación. Se puede responder fácilmente que, en el nivel neuronal, el cerebro humano puede ser muy mecánico y reflexivo, pero que esto no es ningún impedimento para la creatividad y la flexibilidad en el nivel macroscópico. Por supuesto, un opositor podría negar la tesis acerca del mecanismo en el nivel neuronal pero, de cualquier forma, no parece haber ningún buen argumento en favor de la opinión de que la dinámica computacional en un nivel causal básico es incompatible con la creatividad y la flexibilidad en el nivel macroscópico.

Este tipo de objeción puede ganar algún apoyo de la identificación implícita de los sistemas computacionales con los sistemas computacionales *simbólicos*: sistemas que realizan manipulaciones simbólicas de representaciones conceptuales de alto nivel; en el caso extremo, sistemas que inflexiblemente extraen conclusiones a partir de premisas en lógica de primer orden. Tal vez, la objeción tenga alguna fuerza en estos casos, aunque aun esto podría ser discutible. Pero, de cualquier manera, la clase de sistemas computacionales es

mucho más amplia. Una simulación de bajo nivel del cerebro es una computación, por ejemplo, pero no una computación simbólica de esta clase. En un nivel intermedio, los modelos conexionistas de la ciencia cognitiva recurrieron a un tipo de computación que no consiste de manipulación simbólica. En estos casos, puede haber un nivel en el cual el sistema siga reglas, pero ello no se refleja directamente en el nivel de la conducta; precisamente, los conexionistas suelen afirmar que el suyo es un método para producir flexibilidad de alto nivel a partir de mecanicidad de bajo nivel. Como Hofstadter (1979) lo formuló, el nivel en el que pienso no es necesariamente en nivel en el que sumo.[7]

## Objeciones a partir del teorema de Gödel

A veces se sostiene que el teorema de Gödel demuestra que los sistemas computacionales están limitados de un modo que los seres humanos no lo están. El teorema de Gödel nos dice que para cualquier sistema formal consistente lo suficientemente potente como para hacer un cierto tipo de aritmética, existirá un enunciado verdadero —el *enunciado de Gödel* del sistema— que este no puede probar. Sin embargo, nosotros podemos ver que el enunciado de Gödel es verdadero, se argumenta, de modo que tenemos una capacidad de la que el sistema formal carece. Se deduce que ningún sistema formal puede capturar precisamente las capacidades humanas. (Argumentos de este tipo fueron formulados por Lucas [1961] y Penrose [1989, 1994], entre otros.)

La respuesta breve a estos argumentos es que no hay ninguna razón para creer que los seres humanos puedan ver la verdad de los enunciados de Gödel pertinentes. En el mejor de los casos, podemos ver que *si* un sistema es consistente, entonces su enunciado de Gödel es verdadero, pero no hay ninguna razón para creer que podamos determinar la consistencia de sistemas formales arbitrarios.[8] Esto es válido en particular para el caso de sistemas formales complejos, tal como un sistema que simula la salida de un cerebro humano: la tarea de determinar si un sistema de este tipo es consistente bien podría estar más allá de nuestras posibilidades. De modo que podría ocurrir que cada uno de nosotros pueda ser simulado por un sistema formal complejo $F$, tal que no podemos determinar si $F$ es consistente. Si esto es así, no podremos ver la verdad de nuestros propios enunciados de Gödel.

Existen muchas variantes del argumento gödeliano, con respuestas que un opositor podría hacer a esta sugerencia y nuevos caminos secundarios que surgen a su vez. No los analizaré aquí

(aunque lo hago detalladamente en Chalmers 1995c). Estas cuestiones conducen a muchos estimulantes puntos de interés, pero creo que es justo decir que la tesis de que las limitaciones gödelianas no se aplican a los seres humanos nunca se defendió de forma convincente.

## Objeciones a partir de la no computabilidad y la continuidad

Las objeciones de más arriba son argumentos de "alto nivel" de que el funcionamiento cognitivo es no computable. También se podría intentar atacar la posición de la IA en el bajo nivel, argumentando que el funcionamiento físico no es computable. Penrose (1994), por ejemplo, argumenta que podría haber un elemento no computable en una teoría correcta de la gravedad cuántica. Su única evidencia de esta conclusión, sin embargo, se encuentra en el argumento gödeliano de más arriba. No hay nada en la teoría física misma que apoye esa conclusión; de modo que si se voltea el argumento gödeliano, cualquier razón para creer en leyes físicas no computables desaparece. Se podría sostener que dado que todo elemento del cerebro, por ejemplo una neurona, sólo tiene un número finito de estados relevantes, y dado que sólo hay un número finito de elementos relevantes, entonces la estructura causal del cerebro *debe* ser capturable en una descripción computacional.

Esto lleva a la objeción final: que los procesos cerebrales pueden ser esencialmente *continuos* mientras que los procesos computacionales son discretos, y que esta continuidad podría ser esencial para nuestra competencia cognitiva, de modo que ninguna simulación discreta podría duplicarla. Tal vez, al aproximar una neurona por un elemento con sólo un número finito de estados, perdemos algo vital de su funcionamiento. Un oponente podría apelar, por ejemplo, a la presencia de una "dependencia crítica a las condiciones iniciales" en ciertos sistemas no lineales, lo que implica que aun un pequeño error de redondeo en una etapa de procesamiento puede llevar a diferencias macroscópicas importantes en una etapa posterior. Si el procesamiento cerebral es de esta clase, entonces cualquier simulación discreta del cerebro producirá resultados que difieren de la realidad continua.

Hay buenas razones para creer que la continuidad absoluta no puede ser esencial para nuestra competencia cognitiva, sin embargo. La presencia de ruido de fondo en los sistemas biológicos implica que ningún proceso puede depender de requerir más de un cierto nivel de precisión. Más allá de un cierto punto (digamos, el nivel de $10^{-10}$ en una escala apropiada), las fluctuaciones incontrolables en el ruido de

fondo eliminarán cualquier precisión ulterior. Esto significa que si aproximamos el estado del sistema con este nivel de precisión (quizás un poco más allá para estar del lado seguro, por ejemplo, al nivel $10^{-20}$), podríamos tener un desempeño tan fiable como el del propio sistema. Es verdad que debido a los efectos no lineales esta aproximación podría llevar a una conducta diferente de la conducta producida por el sistema en una ocasión dada, pero llevará a conductas que el sistema *podría* haber producido, si el ruido biológico hubiese sido un poco diferente. Podemos incluso aproximar el propio proceso de ruido, si así lo queremos.[9] El resultado será que el sistema de simulación tendrá las mismas *capacidades* conductuales que el sistema original, aunque produzca una conducta específica diferente en ocasiones específicas. La moraleja es que cuando se trata de duplicar nuestras capacidades cognitivas, una aproximación cercana es tan buena como la cosa real.

Es verdad que un sistema con precisión ilimitada podría poseer capacidades cognitivas que ningún sistema discreto podría nunca tener. Por ejemplo, podríamos codificar una cantidad analógica correspondiente al número real cuyo n-ésimo dígito binario es 1 si y sólo si la n-ésima máquina de Turing se detiene con todas las entradas. Utilizando esta cantidad, un sistema continuo perfecto podría resolver el problema de la detención, algo que ningún sistema discreto puede hacer. Pero la presencia de ruido implica que ningún proceso biológico puede implementar fiablemente este sistema. Los sistemas biológicos pueden apoyarse sólo en una cantidad finita de precisión, de modo que el cerebro de los seres humanos y los animales debe estar limitado a capacidades que los sistemas discretos pueden compartir.

## 6. Conclusión

La conclusión general es que no parece haber ninguna barrera de principio para las ambiciones de la inteligencia artificial. Las objeciones externas no parecen tener mucha fuerza. Las objeciones internas podrían ser más preocupantes, pero ninguno de los argumentos en favor de esas objeciones parece convincente cuando se lo analiza. Si los argumentos que enuncié en los capítulos anteriores son correctos, entonces tenemos buenas razones positivas para creer que la implementación de una computación apropiada estará acompañada de la experiencia consciente. De modo que la perspectiva de la conciencia de las máquinas es buena en principio, aunque pueda no serlo todavía en la práctica.

No he dicho gran cosa acerca de exactamente qué *tipo* de computación podría ser suficiente para la experiencia consciente. En

la mayoría de los argumentos utilicé como ejemplo una simulación de neurona por neurona del cerebro; pero es probable que muchos otros tipos de computaciones puedan también ser suficientes. Podría ocurrir, por ejemplo, que una computación que refleje la organización causal del cerebro en un nivel mucho menos detallado pueda todavía capturar lo que es significativo para el surgimiento de la experiencia consciente. Es probable que computaciones de una forma enteramente diferente, que correspondan a tipos totalmente distintos de organización causal, pudiesen también dar origen a ricas experiencias conscientes cuando se las implementa.

Este cuadro es igualmente compatible con los enfoques simbólico y conexionista de la cognición, y también con otros enfoques computacionales. Es posible argumentar que la centralidad de la computación en el estudio de la cognición se deba a que las concepciones computacionales pueden capturar casi *cualquier* tipo de organización causal. Podemos pensar que los formalismos computacionales proporcionan un formalismo ideal para la expresión de patrones de organización causal y ciertamente (en combinación con métodos de implementación) constituyen una herramienta ideal para su reproducción. Cualquiera sea la organización causal que resulte fundamental para la cognición y la conciencia, podemos esperar que una concepción computacional sea capaz de capturarla. Podríamos incluso argumentar que es esta flexibilidad la que subyace a la tan mencionada *universalidad* de los sistemas computacionales. Los defensores de la inteligencia artificial no se encuentran comprometidos con ningún tipo particular de computación como la clase que podría ser suficiente para la mentalidad; la gran plausibilidad de la tesis de la IA se debe precisamente a que la clase de sistemas computacionales es tan amplia.[10]

Es así que la cuestión de exactamente qué clase de computaciones es suficiente para reproducir la mentalidad humana sigue siendo una pregunta abierta; pero tenemos buenas razones para creer que la clase no es vacía.

# 10

# La interpretación de la mecánica cuántica

## 1. Dos misterios

El problema de la mecánica cuántica es casi tan difícil como el problema de la conciencia. La mecánica cuántica nos ofrece un cálculo notablemente exitoso para predecir los resultados de las observaciones empíricas, pero es extraordinariamente difícil comprender la imagen del mundo que supone. ¿Cómo podría nuestro mundo ser del modo que debe ser para que las predicciones de la mecánica cuántica sean exitosas? No hay nada que ni siquiera se aproxime a un consenso en la respuesta a esta pregunta. Al igual que ocurre con la conciencia, con frecuencia parece que *ninguna* solución del problema de la mecánica cuántica puede ser satisfactoria.

Muchas personas pensaron que estos dos problemas tan desconcertantes podrían estar íntimamente vinculados (por ejemplo, Bohm, 1980; Hodgson, 1988; Lockwood, 1989; Penrose, 1989; Squires, 1990; Stapp, 1993; Wigner, 1961). Cuando tenemos dos misterios es tentador suponer que tienen una fuente común. Esta tentación se magnifica por el hecho de que los problemas en la mecánica cuántica parecen estar profundamente vinculados a la noción de observación; esta involucra de modo crucial la relación entre la experiencia de un sujeto y el resto del mundo.

Frecuentemente, se ha sugerido que la mecánica cuántica podría poseer la clave de una explicación física de la conciencia. Pero, como hemos visto, este proyecto nunca llegará a su meta. Al fin de cuentas, las "teorías" cuánticas de la conciencia padecen del mismo tipo de brecha explicativa que las teorías clásicas. En cualquiera de ellas, debe interpretarse la experiencia como algo que va más allá de las

propiedades físicas del mundo. Tal vez la mecánica cuántica pueda desempeñar un papel en la caracterización del vínculo psicofísico, pero la teoría cuántica sola no nos puede decir por qué existe la conciencia.

Pero los problemas pueden estar vinculados de un modo más sutil. Aunque la mecánica cuántica no explica la conciencia, quizás una teoría de la conciencia pueda arrojar luz sobre los problemas de la mecánica cuántica. Después de todo, se acepta en general que estos problemas tienen que ver con la observación y la experiencia. Es natural suponer que una teoría de la experiencia pueda ayudarnos a comprender las cuestiones involucradas. Algunos propusieron un papel activo para la conciencia en la teoría cuántica; sugirieron, por ejemplo, que la conciencia produce el "colapso de la función de onda". Sin embargo, argumentaré en favor de un papel más indirecto para la conciencia en relación con estas cuestiones. En particular, propondré que podemos reconcebir los problemas de la teoría cuántica como problemas acerca de la relación entre la estructura física del mundo y nuestra experiencia de él y que, por consiguiente, una teoría apropiada de la conciencia puede apoyar una interpretación no ortodoxa de la mecánica cuántica.

## 2. El marco conceptual de la mecánica cuántica

El marco conceptual básico de la mecánica cuántica consiste en un cálculo que permite predecir los resultados de las mediciones experimentales. Describiré una versión de ese cálculo aquí, evitando un número de detalles técnicos, con el fin de proporcionar una descripción simple que cubra las características más cruciales. En este apartado, presentaré el marco conceptual tan sólo como un cálculo que permite realizar predicciones empíricas; dejaré abierta la cuestión de si proporciona una descripción directa de la realidad física. Analizaremos los problemas profundos de la interpretación en la siguiente sección.

Dentro de un marco clásico, el estado de un sistema físico puede expresarse en términos muy simples. El estado de una partícula, por ejemplo, se expresa mediante valores determinados de un conjunto de propiedades como la posición y el momento. Podemos denominar a este tipo de valor simple un *valor básico*. Dentro del marco cuántico, las cosas no son tan simples. En general, el estado de un sistema debe expresarse como una *función de onda* o un *vector de estado*. Aquí, las propiedades relevantes no pueden expresarse en valores simples, sino, en cambio, como una especie de combinación de valores básicos. Un estado cuántico puede verse como una *superposición* de estados más simples.

El ejemplo más sencillo es una propiedad como el *spin*, que sólo tiene dos valores básicos.[1] Estos valores básicos pueden rotularse "arriba" y "abajo". En la mecánica cuántica, el spin de una partícula no siempre está arriba o abajo, sin embargo. En general, debe expresarse como una *combinación* de arriba y abajo, cada una con una magnitud compleja diferente. Por lo tanto, es mejor considerar el spin de una partícula como un vector en un espacio vectorial bidimensional. Se lo puede visualizar muy naturalmente como una superposición de un estado de spin arriba y un estado de spin abajo, con diferentes magnitudes para cada uno.

Lo mismo ocurre para la posición y el momento, excepto que cada una de estas tiene un número infinito de valores básicos. La posición y el momento de una partícula clásica pueden tomar cualquiera de un número infinito de valores en un continuo. En correspondencia, la posición de una partícula cuántica debe expresarse en la forma de un vector con un número infinito de dimensiones y una magnitud diferente para cada una de estas localizaciones. Es mejor considerar este vector como una *onda*, con amplitudes diferentes para distintas localizaciones en el espacio; la función que pone en relación una localización con la amplitud correspondiente es la función de onda. De modo similar, el momento de una partícula cuántica puede considerarse como una onda con amplitudes diferentes para los distintos valores básicos del momento. Nuevamente, podemos pensar la posición o el momento de una partícula como una superposición de valores básicos de posición o momento respectivamente.

Debido a que estos estados sólo son vectores, pueden descomponerse en componentes de muchos modos. Aunque suele ser útil considerar un vector de spin bidimensional como una suma de un componente "arriba" y un componente "abajo", se lo puede descomponer de muchos otros modos, según la base elegida para el espacio vectorial. Todas estas bases son igualmente "naturales"; la naturaleza no prefiere ninguna de ellas. De hecho, resulta que un solo vector representa la posición y el momento de una partícula. Si se lo descompone de acuerdo con una base, obtenemos las amplitudes de la "posición"; si se lo descompone de acuerdo con una base diferente, obtenemos las amplitudes del "momento". En general, qué descomposición es relevante en un determinado caso depende de cuál sea la cantidad en la que estamos interesados y, en particular, qué cantidad elegimos *medir*, como expondré en seguida.

Los estados de los sistemas que constan de más de una partícula son algo más complejos, pero la idea básica es la misma. Tómese un sistema que consta de dos partículas, *A* y *B*. El estado del sistema no puede por lo general expresarse mediante una simple combinación

de una función de onda para A y una función de onda para B; los estados de las dos partículas con frecuencia serán *no separables*. Más bien, el estado del sistema debe expresarse como una función de onda en un espacio más complejo. Sin embargo, esta función de onda puede considerarse una especie de superposición de estados más simples del sistema de dos partículas, de modo que el marco general todavía es aplicable. Lo mismo ocurre para sistemas más complejos, en los que un estado todavía se representa mejor como una función de onda que corresponde a una superposición de estados.

Todo esto no es demasiado intuitivo, pero todavía no es paradójico. Si interpretamos este formalismo en su valor nominal como una descripción de la realidad, no es *demasiado* difícil de comprender. Sin embargo, algunos supusieron que esta descripción es incompatible con un enfoque "objetivo" del mundo, ya que implicaría que las entidades no tienen un estado objetivo y determinado. Pero esto no es necesariamente así. Según esta concepción, el estado de una entidad puede expresarse mejor por medio de una función de onda en lugar de mediante cantidades discretas, pero se trata de un estado perfectamente determinado. La concepción simplemente nos dice que, en el nivel básico, la realidad es ondulatoria. Esto requiere un nuevo modo de pensar, pero podemos acostumbrarnos a ello. Después de todo, el nivel básico de la realidad microscópica está muy lejos del nivel macroscópico en el que usualmente nos manejamos y no es del todo sorprendente que tenga propiedades no familiares. Los problemas que puedan surgir provienen de propiedades *ulteriores* de la mecánica cuántica.

El núcleo de la mecánica cuántica consiste en dos principios que determinan la *dinámica* de la función de onda: la *ecuación de Schrödinger* y el *postulado de medición*. Entre ellos, estos dos principios muy diferentes determinan cómo la función de onda de un sistema evoluciona en el tiempo.

La mayor parte de la sustancia de la mecánica cuántica se encuentra en la ecuación de Schrödinger. Esta es una ecuación diferencial que determina cómo evoluciona la función de onda de un sistema bajo *casi* cualquier circunstancia. La estructura detallada de la ecuación no es importante para nuestros propósitos. La característica más significativa aquí es que es una ecuación diferencial *lineal*: dados dos estados A y B tal que A evoluciona en A' y B evoluciona en B', entonces un estado consistente de una superposición de A y B evolucionará en una superposición de A' y B'. También vale la pena hacer notar que bajo la dinámica de la ecuación de Schrödinger, los estados relativamente discretos tienden por lo general a expandirse con el transcurso del tiempo. Un estado que comienza como una

superposición de valores en un dominio limitado por lo general evolucionará en una superposición de valores en un dominio mucho más amplio. Finalmente, esta ecuación es totalmente determinística.

La ecuación de Schrödinger es relativamente simple y bien comprendida. Es aquí donde reside lo más básico de la teoría cuántica. Al aplicar la teoría cuántica a un problema práctico o experimental, el grueso del trabajo consiste en calcular cómo evolucionan los diversos estados según la dinámica de Schrödinger.

Sin embargo, la ecuación de Schrödinger no puede ser *todo* lo que hay para decir. Según la ecuación, la vasta mayoría de los estados físicos pronto evolucionará en una superposición de un amplio dominio de estados. Pero esto no concuerda con nuestras observaciones del mundo. Cuando medimos la posición de una partícula, encontramos un valor definido, no la superposición de valores que la ecuación de Schrödinger predice. Si la ecuación de Schrödinger fuese todo lo que hay en la dinámica cuántica, entonces aun en el nivel macroscópico el mundo evolucionaría en un estado ampliamente superpuesto. Sin embargo, en nuestra experiencia no lo hace. Los indicadores tienen localizaciones definidas, los objetos en movimiento tienen un momento medible definido, etc. De modo que debe haber más en la historia: algo que nos lleve de la ecuación a los tipos de sucesos discretos que caracterizan nuestra experiencia.

La segunda parte de la historia en el formalismo estándar es el *postulado de medición* (también conocido como el postulado de colapso o proyección). Este afirma que bajo circunstancias especiales la dinámica de Schrödinger no se aplica. Específicamente, dice que cuando se realiza una *medición*, la función de onda *colapsa* en una forma más definida. El modo como colapsa depende de la propiedad que se mida. Por ejemplo, si medimos el spin de una partícula, aunque con anterioridad esté en un estado superpuesto, colapsará en un estado en el cual el spin es arriba o abajo. Si medimos la posición de una partícula, su función de onda colapsará en un estado con una posición definida.[2] El estado resultante todavía corresponde a una función de onda, pero es una función de onda en la cual toda la amplitud está concentrada en una posición determinada; la amplitud en cualquier otra posición es cero. A cualquier cantidad que podamos medir le corresponde un operador; en la medición el estado colapsará en un *autoestado* de ese operador. Un autoestado de un operador es siempre un estado en el cual la cantidad medible correspondiente tiene un valor definido. Se deduce entonces que cuando realizamos una medición de una cantidad, siempre resulta un valor definido de esa cantidad, lo que concuerda precisamente con nuestra experiencia.

La dinámica del colapso es probabilística, no determinística. Si una partícula está en un estado que es una superposición de posiciones, entonces cuando se mide la posición sabemos que colapsará en un estado con una posición definida, pero no sabemos cuál será esa posición. Más bien, para cada estado colapsado potencial, el postulado de medición especifica la *probabilidad* de que el sistema colapse en ese estado. Esta probabilidad[3] está dada por el cuadrado de la amplitud de la función de onda en el lugar correspondiente al valor definido en cuestión. Por ejemplo, si el spin de una partícula es una superposición de spin arriba (con amplitud $\frac{1}{2}$) y spin abajo (con amplitud $\frac{\sqrt{3}}{2}$) entonces, cuando se mide el spin, este colapsará en un estado de spin arriba con probabilidad de $\frac{1}{4}$ o en un estado de spin abajo con probabilidad de $\frac{3}{4}$. Las amplitudes en una función de onda siempre tienen la propiedad de que las probabilidades correspondientes suman 1.

## 3. La interpretación de la mecánica cuántica

En conjunto, estos dos principios constituyen un cálculo extremadamente poderoso para predecir los resultados de las mediciones experimentales. Para predecir los resultados de un experimento, expresamos el estado de un sistema como una función de onda y calculamos cómo esta evoluciona en el tiempo de acuerdo con la ecuación de Schrödinger, hasta el punto en que se realiza una medición. En ese momento, utilizamos las amplitudes de la función de onda calculada para determinar la probabilidad con que resultarán los diversos estados colapsados y calcular la probabilidad de que la medición produzca cualquier cantidad determinada. Los resultados experimentales prestaron constantemente su apoyo a las predicciones de la teoría; pocas teorías científicas han sido tan exitosas en su tarea predictiva. Como cálculo, la teoría casi no tiene fallas.

Los problemas surgen cuando preguntamos *cómo* puede ser que el cálculo funcione. ¿Qué podría estar ocurriendo en el mundo real para hacer que las predicciones del cálculo sean tan precisas? Este es el problema de la *interpretación* de la mecánica cuántica. Existen muchas diferentes opciones disponibles para enfocar este problema, ninguna de las cuales es totalmente satisfactoria.

### Opción 1: Tomar el cálculo literalmente

La primera reacción natural es tomar el formalismo de la mecánica cuántica en su valor nominal, como lo hacemos con la mayoría de las teorías científicas. El cálculo involucra una función de onda gobernada por la dinámica de la ecuación de Schrödinger y el postu-

lado de medición y, además, funciona, de modo que deberíamos suponer que nos da una imagen directa de qué es lo que ocurre en el mundo. Es decir, el estado de un sistema en la realidad es precisamente el estado de onda expresado por la función de onda que evoluciona de acuerdo con la dinámica expresada por los dos principios básicos. La mayor parte del tiempo, el estado evoluciona de acuerdo con la ecuación de Schrödinger, pero cuando se hace una medición evoluciona de acuerdo con el postulado de medición. Según este enfoque, el mundo consiste de ondas que usualmente evolucionan en forma lineal en una superposición y que ocasionalmente colapsan en un estado más definido cuando se realiza una medición.

Pero no es fácil comprender esta imagen. Todos los problemas surgen del postulado de medición. Según este postulado, ocurre un colapso cuando se realiza una medición, pero ¿qué es una medición? ¿Cómo sabe la *naturaleza* cuándo se hace una medición? "Medición" no es, seguramente, un término básico en las leyes de la naturaleza; para que el postulado de medición pueda ser al menos remotamente plausible como ley fundamental, la noción de medición deberá ser reemplazada por algo más claro y más básico. Si el colapso de una función de onda es un proceso objetivamente existente en el mundo, entonces necesitamos criterios objetivos claros de cuándo esto ocurre.

Una solución obviamente insatisfactoria es decir que un colapso ocurre siempre que un sistema cuántico interactúa con un *aparato de medición*. El problema es que resulta muy poco verosímil que la noción "aparato de medición" aparezca en las leyes básicas, tal como sucede con la noción de "medición". Antes, necesitábamos criterios para determinar qué podía considerarse una medición; ahora necesitamos criterios para determinar qué es un aparato de medición.

Una sugerencia popular en los días iniciales de la mecánica cuántica era que un aparato de medición es un sistema *clásico*, y que una medición ocurre cada vez que un sistema cuántico interactúa con un sistema clásico. Pero esto es claramente insatisfactorio. Se supone que la teoría cuántica es una teoría universal y que debería aplicarse a los procesos dentro de un instrumento de medición tanto como se aplica a los procesos en cualquier otro lado. A menos que supongamos que hay dos tipos fundamentalmente diferentes de objetos físicos en el mundo —una suposición que requeriría el desarrollo de una teoría enteramente nueva—, entonces el "sistema clásico" no puede ser un término en una ley fundamental de la naturaleza, no más de lo que puede serlo una "medición".

Una sugerencia relacionada es que una medición ocurre cada vez que un sistema cuántico interactúa con un sistema *macroscópico*. Pero es igualmente claro que "macroscópico" no es una noción que

pueda figurar en una ley básica. Debe ser reemplazado por algo más preciso: algo como "sistema con masa de un gramo o mayor". Pero, sería extraordinariamente arbitrario que algo así figurase en una ley básica.

No existe ningún criterio físico del colapso que parezca ni remotamente aceptable. Un criterio formulado en el nivel microscópico —que sugiera que el colapso ocurre cuando un sistema interactúa con un protón, por ejemplo— queda descartado por los resultados experimentales. La alternativa es que el criterio deba involucrar una propiedad física de nivel superior, de modo que el colapso ocurra cuando los sistemas adoptan una cierta configuración de alto nivel. Pero cualquier propiedad semejante de alto nivel parecería arbitraria, y nunca se propuso ningún candidato plausible. También, hay algo muy extraño en el supuesto de que la dinámica de Schrödinger de los sistemas microscópicos deba repentinamente anularse cuando esos sistemas se encuentran en el contexto de ciertas configuraciones especiales.

El único criterio remotamente sostenible propuesto es que una medición ocurre cuando un sistema cuántico afecta la *conciencia* de algún ser. A diferencia de los criterios previos, este criterio está al menos determinado y no es arbitrario.[4] La interpretación correspondiente del cálculo es razonablemente elegante y simple en su forma, y es la única interpretación *literal* del cálculo que tiene alguna aceptación general. Esta interpretación fue sugerida por primera vez por London y Bauer (1939), pero está más estrechamente asociada con Wigner (1961).

Nótese que esta interpretación *presupone* un dualismo mente-cuerpo. Si la conciencia fuese sólo otra propiedad física, entonces surgirían todos los problemas previos. La propuesta se volvería otro enfoque de "propiedad de alto nivel", en el cual las funciones de onda de los sistemas físicos colapsan en el contexto de ciertas configuraciones físicas complejas. Por otro lado, si el dualismo es válido, entonces el criterio del colapso puede ser verdaderamente fundamental. Más aún, el hecho de que la causa del colapso sea externa al procesamiento físico hace posible una teoría mucho más simple. Todos los sistemas puramente físicos estarían ahora gobernados únicamente por la dinámica de Schrödinger; la muy diferente dinámica de la medición tendría una fuente independiente.

Sin embargo, esta interpretación tiene algunas consecuencias contrarias a la intuición. Tómese un aparato de medición como un indicador que mide el estado de un electrón y supóngase que el estado de este último inicialmente está superpuesto. Si no hay conciencia en la vecindad, todo el sistema estará gobernado por la dinámica lineal

de Schrödinger: dado que diferentes estados discretos del electrón producirían diferentes estados discretos del indicador, se deduce que un estado superpuesto del electrón producirá un estado *superpuesto* del indicador. Es decir, ¡la teoría predice que el indicador apunta a muchas localizaciones diferentes simultáneamente! Sólo cuando *miro* el indicador este apunta a una posición definida.

El escenario del gato de Schrödinger tiene consecuencias aun más extrañas. En este escenario un gato está encerrado dentro de un gabinete; mediante un instrumento se mide el spin de un electrón y hay un aparato montado de modo que mate al gato si y sólo si el spin del electrón está "arriba". (Asúmase que el gato está anestesiado, de modo que su conciencia no entra en la situación.) ¡Si el electrón está inicialmente en un estado superpuesto, entonces el gato se moverá a un estado que es una superposición de la vida y la muerte! Sólo cuando un ser consciente mira dentro del gabinete el estado de vida o de muerte del gato quedará determinado.

En esta cuadro, *cualquier* sistema macroscópico estará usualmente en una superposición a gran escala si no hay conciencia en la vecindad. Antes de que la conciencia hubiese evolucionado, todo el universo estaba en una gigantesca superposición, hasta que supuestamente la primera mota de conciencia causó que su estado repentinamente colapsase. Esto puede sonar descabellado, pero es una consecuencia directa de la única interpretación literal defendible de los principios de la mecánica cuántica. Espero que esto ayude a poner de relieve lo extraña que es la mecánica cuántica y la gravedad de los problemas planteados por su interpretación.

Las consecuencias contrarias a la intuición tal vez podrían aceptarse, pero, no obstante, yo no defiendo esta interpretación. Para empezar, es incompatible con el enfoque que sostuve según el cual la conciencia es ubicua. Si la conciencia está asociada incluso con los sistemas muy simples, entonces, según esta interpretación, el colapso sucederá en todo nivel básico y muy frecuentemente. Esto es inconsistente con la evidencia física, que requiere que las superposiciones de bajo nivel pemanezcan por lo general no colapsadas durante un tiempo significativo. Un segundo problema es que no hay nada que se aproxime a una buena teoría acerca de qué tipo de efecto sobre la conciencia tiene el colapso o acerca de qué forma adoptará este último. Hay muchos modos diferentes en los que esto podría especificarse, pero ninguna única forma de especificar los detalles parece particularmente convincente.

Otros problemas surgen de la propia noción de colapso. Para empezar, este debe ser *no local*: cuando dos partículas tienen estados entremezclados, la medición de la primera partícula causará que el

estado de la segunda colapse simultáneamente. Esto lleva a cierta tensión con la teoría de la relatividad. Por ejemplo, parece que el colapso no local requiere recurrir a un marco de referencia privilegiado. Sin un marco de referencia de este tipo, el tiempo de colapso de la segunda partícula estará subdeterminado, ya que la simultaneidad en diferentes localizaciones no está bien definida.

Más en general, todo el proceso del colapso no encaja bien con el resto de la física. Tomado literalmente, es un proceso no local, temporalmente asímétrico, discontinuo e instantáneo que es totalmente diferente de cualquier otro proceso que la teoría física postula. Parece extraño que un proceso tan curioso deba existir junto con la ecuación de Schrödinger, que es simple, local, temporalmente simétrica y continua. En comparación con la elegancia y la potencia de la ecuación de Shrödinger, que está en el centro de la teoría cuántica, el colapso parece casi un elemento arbitrario y agregado. Existe algo muy extraño en la idea de que el mundo tiene dos tipos totalmente diferentes de dinámica en su nivel básico.

Estos están lejos de ser argumentos contundentes, por supuesto, y la interpretación en la que la conciencia colapsa la función de onda merece ser tomada muy en serio. Sin embargo, creo que hay buenas razones para buscar otra interpretación, una que nos dé un punto de vista más simple y directo de los procesos básicos de la naturaleza.

### Opción 2: Intentar obtener el postulado de medición en forma gratuita

Todos los problemas con la interpretación literal surgen de tomar el postulado de medición como ley fundamental. Es tentador suponer, en cambio, que el postulado podría ser *no básico*, una consecuencia de principios más fundamentales. Existen dos formas en las que esto podría ocurrir. Podríamos tratar de introducir principios básicos *ulteriores*, menos problemáticos que el postulado de medición, pero que tengan el mismo efecto. Esta es la estrategia de la opción 4. O podríamos tratar de derivar los efectos como una consecuencia de principios básicos conocidos, tal como la ecuación de Schrödinger. Esto es, podríamos tratar de obtener el postulado de medición en forma gratuita.

Es fácil ver la motivación intuitiva de esta estrategia. Tenemos la intuición de que los efectos de superposición se aplican principalmente en un nivel microscópico y podrían de algún modo "cancelarse" en el nivel macroscópico. Tal vez, cuando hay muchas superposiciones microscópicas, estas interactúan de modo de producir un estado macroscópico que está relativamente definido. Debido a algunas

propiedades matemáticas de las configuraciones complejas, podríamos ver cómo un colapso *efectivo* podría ser consecuencia de la indefinición microscópica. Un colapso probabilístico fundamental sería entonces reemplazado por un proceso estadístico emergente en un sistema complejo.

Han existido numerosos intentos de desarrollar la matemática involucrada, con frecuencia apelando a los principios estadísticos de la termodinámica (por ejemplo, Daneri, Linger y Prosperi, 1962). Desafortunadamente, todos estos intentos fallaron, y en la actualidad se acepta que *deben* fallar. Debido a que la dinámica de Schrödinger es lineal, siempre es posible construir situaciones en las cuales las superposiciones microscópicas llevan a superposiciones macroscópicas. Si un electrón "arriba" lleva a un estado macroscópico, y un electrón "abajo" lleva a otro, entonces un electrón superpuesto debe llevar a un estado macroscópico superpuesto (Albert, 1992, p. 75, ofrece un argumento muy directo sobre este punto). A menos que se introduzcan nuevos principios básicos, debemos esperar superposiciones en el nivel macroscópico.

Estas estrategias tienen algo para ofrecer. Este tipo de recurso a la estadística, como también los trabajos más recientes sobre la "descohesión" de Gell-Mann y Hartle (1990) y otros, sugiere que una función de onda superpuesta con frecuencia se resolverá en una superposición relativamente bien definida de estados macroscópicos distintos, en lugar de ser una confusión. Estos estados macroscópicos se "descohesionan" unos de los otros, con sólo mínimos efectos de interferencia entre ellos. Esto al menos nos ayuda a encontrar algún elemento del mundo clásico familiar en la función de onda superpuesta. Pero la función de onda es todavía una superposición y nada en este tipo de trabajos nos dice por qué sólo un elemento de la superposición macroscópica debería ser real. De modo que se necesitan nuevos trabajos para resolver el problema básico. Este tipo de enfoque es quizá más útil cuando se lo combina con una de las otras opciones, como la opción 5.

## Opción 3: De lo cual no podemos hablar...

Tal vez el enfoque dominante entre los físicos actuales sea que simplemente no deberíamos preguntar qué ocurre en el mundo real detrás del cálculo de la mecánica cuántica. El cálculo funciona y eso es todo. Existen dos versiones de este enfoque. De acuerdo con la primera versión, tal vez *algo* ocurra en el mundo, pero nunca podremos saber qué es. El cálculo nos da toda la información empírica que podremos alguna vez llegar a tener, de modo que cualquier cosa

ulterior es pura especulación. De este modo, es mejor dejar de preocuparnos y continuar calculando. Este enfoque tiene sentido para los propósitos prácticos, pero es insatisfactorio para cualquiera que desee que la física nos hable acerca del nivel fundamental de la realidad. Ya que el cálculo funciona, queremos tener al menos alguna idea de *cómo* es que lo hace. Tal vez nunca podamos saberlo con seguridad, pero tiene sentido preguntar.

La segunda versión adopta una postura más dura; sostiene que no hay ningún hecho empírico en lo que ocurre en el mundo. Según este enfoque, los hechos se agotan en el hecho de que el cálculo funciona. Este punto de vista no suele enunciarse tan explícitamente, tal vez debido a que formulado de esta manera el enfoque es casi imposible de creer. ¡Nos ofrece una imagen de la realidad que deja afuera al mundo! Lleva a una versión del idealismo, en la cual todo lo que existe son nuestras percepciones o algo muy cercano a esto. Antes de abrir el gabinete que contiene al gato de Schrödinger, este no está en un estado muerto, no está en un estado vivo y no está en un estado superpuesto; simplemente no está en ningún estado. Al desistir de que es un hecho empírico lo que encontramos detrás de nuestras mediciones, este enfoque desiste de una realidad independientemente existente.

La "interpretación de Copenhague" formulada por Bohr y sus colegas suele interpretarse como una versión de este enfoque, aunque los escritos de Bohr son algo ambiguos y su interpretación no es fácil. Estos escritos también sugieren a veces elementos de la primera opción y de la versión epistemológica de esta opción. Bohr puso un gran énfasis sobre la naturaleza "clásica" de un aparato de medición, y puede interpretarse que sus puntos de vista sugieren que sólo los objetos clásicos (o macroscópicos) tienen un estado objetivo. Las cuestiones acerca del estado real de un objeto descripto por una superposición simplemente están proscritos. Pero esto se basa en una división entre los sistemas clásicos y cuánticos que es difícil de trazar sobre la base de criterios objetivos; es difícil de imaginar que la realidad simplemente se "desvanezca" cuando descendemos desde el nivel macroscópico al microscópico. Muchos creen que si se toma en serio el punto de vista de Bohr, este lleva al operacionalismo fuerte que mencionamos en el último párrafo. Al igual que esa perspectiva, ofrece una imagen del nivel básico de la realidad que no es ninguna imagen en absoluto.

### Opción 4: Postular principios físicos básicos ulteriores

Dado que la interpretación literal del postulado de medición es inaceptable, y de que no puede derivarse de los principios físicos existentes, es natural suponer que algo más debe estar ocurriendo. Tal vez, si postulamos principios físicos básicos *ulteriores* podamos ser capaces de explicar la efectividad del cálculo de la mecánica cuántica de un modo menos problemático.

La primera manera de hacer esto es conservar la idea del colapso, pero explicarla de modo diferente. Una estrategia de este tipo conserva el supuesto de que los estados físicos básicos son funciones de onda gobernadas por la ecuación de Schrödinger, pero introduce nuevos principios para explicar cómo las superposiciones microscópicas se transforman en fenómenos macroscópicos discretos.

El ejemplo más conocido de esta estrategia es la interpretación "GRW" debida a Ghirardi, Rimini y Weber (1986; véase también Bell, 1987a).[5] Esta interpretación postula una ley fundamental según la cual el vector de estado de posición de cualquier partícula elemental puede sufrir un "colapso" microscópico en cualquier momento, con alguna probabilidad muy pequeña (la probabilidad de que una partícula colapse en un segundo determinado es aproximadamente de uno en $10^{15}$). Cuando ocurre un colapso de este tipo, por lo general llevará a un colapso del estado de un sistema macroscópico en el que está inmerso, debido a los efectos de no separabilidad. Existen muchas partículas de este tipo en cualquier sistema macroscópico, de manera que se deduce que cualquier sistema macroscópico dado en cualquier momento particular usualmente estará en un estado relativamente discreto. Es posible mostrar que esto se acerca mucho a reproducir las predicciones del postulado de medición.

La alternativa es eliminar la necesidad del colapso negando que el nivel básico de la realidad esté representado por una función de onda superpuesta. Si propiedades como la posición tienen valores determinados incluso en el nivel básico, entonces el colapso no tiene por qué ocurrir nunca. Una teoría de este tipo postula "variables ocultas" en el nivel básico, lo que explica directamente la discreticidad de la realidad en el nivel macroscópico. El costo de esta sugerencia es que se necesitan nuevos principios para explicar por qué los principios de la evolución y el colapso de la función de onda *parecen* funcionar tan bien.

El ejemplo más destacado aquí es la teoría desarrollada por Bohm (1952). Según ella, la posición de las partículas básicas está siempre determinada. La función de onda conserva el papel de una especie de "onda piloto" que guía la evolución de la posición de una partícula; la

función de onda está gobernada por la ecuación de Schrödinger. Las predicciones probabilísticas del postulado de medición se interpretan como *leyes estadísticas*. Según esta teoría nunca podemos saber la posición exacta de una partícula antes de medirla, sólo su función de onda. El postulado de medición nos dice la *proporción* de partículas con una función de onda dada que tendrán una posición determinada. Por lo tanto produce las mejores predicciones estadísticas que podemos esperar, dada nuestra ignorancia.

Todas las propuestas de esta clase tienen problemas. Tanto la interpretación GRW como la interpretación de Bohm le asignan una especial determinación a la *posición*, lo que quiebra así la simetría entre posición y momento en el cálculo de la mecánica cuántica. Esto tiene sentido para propósitos predictivos, ya que puede suponerse que posiciones determinadas siempre subyacen a nuestros juicios de determinación macroscópica (piénsese en la posición de un indicador, por ejemplo), pero contribuye a una teoría más excéntrica. Por razones vinculadas, existen serias dificultades para reconciliar estos enfoques con la teoría de la relatividad.

La teoría GRW tiene algunas otras dificultades, quizá la más seria de las cuales es que no implica estrictamente que el mundo macroscópico sea discreto en absoluto. Un estado macroscópico todavía se representa mediante una función de onda superpuesta; aunque la mayor parte de su amplitud está concentrada en un lugar, la amplitud es distinta de cero dondequiera que la amplitud de la función de onda no colapsada sea distinta de cero. De modo que los problemas de la superposición se repiten. El indicador todavía apunta a muchas localizaciones, aun después de una medición. Es verdad que la amplitud para la mayoría de estas localizaciones es muy pequeña, pero es difícil justificar cómo una superposición de baja amplitud podría ser más aceptable que una de alta amplitud.

La teoría de Bohm tiene menos problemas técnicos que la interpretación de GRW, pero posee algunas consecuencias extrañas. Notablemente, es *no local* en un grado extraordinario. (Cualquier teoría de variables ocultas que satisfaga las predicciones del cálculo debe ser no local, por las razones enunciadas en Bell, 1964.)[6] No sólo se trata de que las propiedades de una partícula pueden afectar las propiedades de otra partícula a cierta distancia instantáneamente. ¡También ocurre que para determinar la trayectoria de una partícula, podríamos vernos obligados a tomar en cuenta las funciones de onda de partículas en otras galaxias! Todas estas cosas tienen un papel en la composición de la función de onda global, y esa función de onda gobierna simultáneamente las trayectorias de las partículas en todo el universo.

Tal vez la razón más importante para sospechar de estas interpretaciones, sin embargo, es que postulan *complejidad detrás de la simplicidad*. Cualesquiera sean sus problemas, el cálculo de la mecánica cuántica es extraordinariamente simple y elegante. Estas interpretaciones, por otro lado, introducen principios ulteriores *ad hoc* para reemplazar y explicar ese marco conceptual simple. Esto se aplica ligeramente menos a la interpretación GRW, cuya complejidad ulterior sólo consiste en introducir dos nuevas constantes fundamentales y en romper la simetría entre posición y momento; pero, sigue siendo el caso que es extraordinariamente "afortunado" que los valores de las constantes sean tales como para casi reproducir las predicciones del marco estándar. La complejidad extra de la interpretación de Bohm es peor: postula posiciones determinadas *y* una función de onda, un principio fundamental nuevo y complejo por el cual la función de onda determina la posición de las partículas, y quiebra la simetría del marco original.

Podríamos decir que estas interpretaciones hacen parecer como si el mundo hubiese sido construido por el demonio maligno de Descartes, ya que nos llevan a creer que el mundo es de un modo cuando en realidad es de otro. Como Albert y Loewer (1989) lo formulan, el Dios del enfoque de Bohm no juega a los dados, pero tiene un malicioso sentido del humor. El escenario en el cual la interpretación compleja de Bohm duplica las predicciones del marco simple sólo difiere en grado del caso en el cual las entradas a un cerebro en un tanque son manipuladas para producir la apariencia de un mundo externo directo. Es reminiscente de una "interpretación" de la teoría evolutiva de acuerdo con la cual Dios creó el registro fósil intacto hace unos pocos miles de años y se aseguró de que las predicciones de la teoría evolutiva se duplicasen. La simplicidad de un marco explicativo ha sido sacrificada por una hipótesis compleja que reproduce los resultados de la teoría original.

El marco conceptual de la mecánica cuántica es tan simple y elegante que una teoría básica que no reproduzca esa simplicidad y elegancia nunca podrá ser satisfactoria o totalmente plausible. Si existiesen unas pocas anomalías en la teoría cuántica, algunos resultados experimentales que el marco no predijese perfectamente, podría ser más plausible pensar que esa simplicidad es la punta de un complejo témpano. En su estado actual, sin embargo, el marco es tan sólido que parece extraordinario que tengamos que postular un aparato complejo para explicar sus simples predicciones.

Dados los problemas que *todas* las interpretaciones de la mecánica cuántica poseen, todas ellas deben tomarse seriamente. Pero es natural que busquemos una imagen más simple del mundo.

### Opción 5: La ecuación de Schrödinger y nada más

El eje de la mecánica cuántica es la ecuación de Schrödinger, que está presente de una forma u otra en todas sus interpretaciones. Todas las interpretaciones que hemos considerado agregan alguna cosa a la ecuación de Schrödinger para explicar la discreticidad macroscópica del mundo. Pero, de lejos, la más simple es la que dice que la ecuación de Schrödinger es válida y nada más. Es decir, el estado físico del mundo está completamente descripto por una función de onda, y su evolución está completamente descripta por la ecuación de Schrödinger. Esta es la interpretación dada por Everett (1957, 1973).

Una estrategia examinada anteriormente (opción 2) también sostenía que la ecuación de Schrödinger era todo, pero argumentaba que esto es compatible con la discreticidad en el nivel macroscópico. Vimos que esto debía fallar por razones matemáticas simples. La interpretación de Everett es mucho más radical. Según este enfoque, debe tomarse la ecuación de Schrödinger por su valor nominal, y el estado del mundo en cada nivel se describe mediante una función de onda. Se deduce que, en contra de las apariencias, el mundo está en un estado superpuesto aun en el nivel macroscópico.

## 4. La interpretación de Everett

La motivación para esta interpretación es obvia. El corazón de la mecánica cuántica es la ecuación de Schrödinger. El postulado de medición y todos los otros principios que fueron propuestos parecen extras agregados. ¿Entonces por qué no deshacerse de ellos? El problema con esta interpretación es igualmente obvio. Si la ecuación de Schrödinger es todo, entonces el mundo está superpuesto en todo nivel. Pero no *parece* superpuesto: nunca percibimos indicadores que estén en una superposición de dos estados. ¿Por qué no?

En el mejor de los casos, esta interpretación es sumamente contraria a la intuición. De acuerdo con este enfoque, no sólo el estado del electrón puede describirse mejor mediante una superposición, ¡sino también el estado de un indicador que lo mide! Objetivamente, no es del todo cierto decir que el indicador apunta hacia arriba o apunta hacia abajo. Más bien, está en una superposición de los estados de señalar hacia arriba y hacia abajo. Lo mismo ocurre con el estado macroscópico de casi todo: está en un estado que puede describirse mediante una función de onda que casi nunca corresponderá a un solo estado "discreto". La superposición, según este punto de vista, está en todos lados. ¿Por qué entonces el mundo parece discreto?

La respuesta de Everett a esta pregunta es *extender la superposición todo el camino hasta la mente.* Si tomamos en serio la ecuación de Schrödinger, entonces si el indicador que mide un electrón está en una superposición de estados, el cerebro de una persona que percibe el indicador estará él mismo en una superposición. Se describirá el estado del cerebro como una superposición de un estado en el cual percibe que el indicador apunta hacia arriba y otro estado en el cual percibe que el indicador apunta hacia abajo. El paso clave de Everett es suponer que cada uno de esos estados debería estar asociado a un observador separado. Lo que ocurre luego de una medición es que se producen dos observadores. Uno de ellos experimenta un indicador "arriba" y el otro percibe un indicador "abajo". Se deduce que *cada* observador experimentará un estado discreto del mundo.

Everett muestra que según este marco conceptual, los observadores tendrán la mayor parte de las propiedades que esperamos que los observadores tengan, y que la mayoría de las predicciones del cálculo de la mecánica cuántica pueden derivarse. Por ejemplo, no es difícil ver que cada uno de los dos estados superpuestos no tendrá ningún acceso al otro estado superpuesto, de modo que la superposición de la mente no se revelará en ningún único estado. Incluso es posible mostrar que cuando un observador que hace una medición percibe a otro observador que mide la misma cantidad, el resultado percibido de las mediciones concordará, de modo que el mundo parecerá bastante coherente. En síntesis, cualquier único observador experimentará el mundo básicamente del modo que esperamos, aun cuando este se encuentre en un estado superpuesto.

Esta interpretación no debería confundirse con la interpretación del *desdoblamiento de mundos*, según la cual el mundo literalmente se divide en muchos mundo separados cada vez que se realiza una medición. Existe un mundo en el cual el indicador señala hacia arriba, y un mundo totalmente independiente en el cual el indicador señala hacia abajo. Tomado de esta forma, el enfoque está muy lejos de ser simple. Para empezar, se requiere un nuevo y extraordinario principio básico para describir el proceso de "desdoblamiento". Está lejos de ser evidente cuándo exactamente debería ocurrir el "desdoblamiento" (el problema de la "medición" revivido en una nueva forma) y es muy poco evidente cuáles deberían ser los mundos resultantes de un desdoblamiento. Para que ocurra un desdoblamiento literal, debe "desdoblarse" la función de onda en numerosos componentes; pero hay muchos modos de descomponer una función de onda y la mecánica cuántica no produce ninguna base preferida para la descomposición. Esta interpretación parece aun más compleja y *ad hoc* que las diversas interpretaciones del "colapso", y hay pocas razones para aceptarla.

El enfoque del desdoblamiento suele atribuirse a Everett (principalmente debido a las exposiciones de su trabajo realizadas por DeWitt (1970, 1971), pero es difícil encontrarlo en sus publicaciones. El punto de vista de Everett no es del todo claro, pero puede interpretárselo en forma mucho más natural del modo que sugerí; esta interpretación es también recomendada por Albert y Loewer (1988) y Lockwood (1989). Según esta perspectiva, no existe ningún "desdoblamiento" objetivo. Más bien, la función de onda evoluciona en una superposición de estados que conviene considerar como componentes de un único mundo. El enfoque de Everett se denomina a veces una interpretación de *muchos mundos* (lo que sugiere el enfoque de desdoblamiento de mundos), pero la perspectiva que analizo es más precisamente una interpretación de *único gran mundo*. Sólo hay un mundo, pero hay más en él de lo que hubiéramos pensado.[7]

Según este enfoque, si existe algún desdoblamiento, es sólo en la mente de los observadores. Cada vez que las superposiciones afectan el estado cerebral de un sujeto, resulta un número de mentes separadas que corresponden a los componentes de la superposición. Cada una de estas percibe un mundo discreto separado que corresponde al tipo de mundo que percibimos, llamémoslo un *minimundo*, en oposición al *maximundo* de la superposición. El mundo real es un maximundo y los minimundos sólo están en la mente de los sujetos. Everett denomina a este enfoque una interpretación de *estados relativos*: el estado de un minimundo, en el cual los indicadores apuntan a posiciones discretas, sólo representa el estado del mundo *relativo* a la especificación de un observador. El estado objetivo del mundo es una superposición.

Sin embargo, un elemento clave es dejado sin analizar en esta interpretación. ¿Por qué es legítimo identificar cada componente de un estado cerebral asociado con un observador distinto? ¿Por qué no hay, en cambio, un solo observador con un estado mental superpuesto y confuso? ¿Por qué un estado cerebral incoherente no da origen a ninguna mente en absoluto? El tratamiento de Everett evade estas preguntas cruciales. Podría parecer que al asociar la función de onda de un estado cerebral con un número de mentes donde cada una percibe un estado discreto, Everett hace una apelación ilegítima a una *base preferida*, así como lo hacía la interpretación del desdoblamiento de mundos. Una función de onda no viene con una división objetiva en componentes, sino que puede descomponerse de muchos modos, dependiendo de la elección de una base para el espacio vectorial correspondiente. Con frecuencia es natural para nuestros propósitos descomponer una función de onda de algún modo, según una base particular, pero una descomposición de esta clase no refleja

una propiedad objetiva de la función de onda. Aunque el estado cerebral puede descomponerse en un estado "perceptor arriba" y uno "perceptor abajo", también puede descomponerse en dos estados cada uno de los cuales tiene percepciones confundidas. Al postular una descomposición objetiva, Everett parece ir más allá de los recursos que la ecuación de Schrödinger proporciona.

El elemento crucial omitido del tratamiento de Everett es un análisis de la relación entre la mente y el cuerpo. Everett supone que un estado cerebral superpuesto tendrá un número de sujetos distintos de experiencia asociados con él, pero no hace nada por justificar ese supuesto. Es evidente que esta cuestión depende crucialmente de una teoría de la conciencia. Penrose (1989) hace una sugerencia similar:

> En particular, no veo por qué un ser consciente debe percatarse de sólo "una" de las alternativas en una superposición lineal. ¿Qué es lo que en la conciencia exige que uno no pueda "percatarse" de esa exasperante combinación lineal de un gato muerto y un gato vivo? Me parece que se necesitaría una teoría de la conciencia antes de que el enfoque de muchos mundos pueda conciliarse con lo que realmente observamos. (p. 296)

Es posible interpretar la cuestión central en la mecánica cuántica como un problema de la relación entre los procesos físicos y la experiencia. El eje de la mecánica cuántica es la imagen en la cual la realidad microscópica se describe mediante una función de onda superpuesta que evoluciona de acuerdo con la ecuación de Schrödinger. Pero, *experimentamos* el mundo como discreto. La pregunta fundamental es cómo sucede esto. Diferentes interpretaciones ofrecen diferentes respuestas. Algunas (como la de Bohm) niegan la primera premisa: afirman que la realidad es discreta aun en el nivel básico. Algunos plantean principios básicos (el postulado de medición o la ley de colapso de GRW) para realizar una transición desde lo superpuesto a lo discreto. Algunas teorías (las de la opción 2) intentan explicar cómo los estados microscópicos superpuestos pueden producir estadísticamente una realidad macroscópica discreta. Estas últimas tres estrategias son todas estrategias *indirectas* que intentan explicar la discreticidad de la experiencia explicando una discreticidad subyacente a la realidad macroscópica.

Una estrategia alternativa es responder la pregunta acerca de la experiencia *directamente*. Si tomamos en serio la supremacía de la ecuación de Schrödinger, la pregunta fundamental será por qué, dado que la estructura física del mundo es como *esto*, lo experimenta-

mos como *aquello*. Esta es precisamente una pregunta acerca del modo como ciertas estructuras físicas dan origen a la experiencia. Es decir, es el tipo de pregunta que he analizando a través de todo libro, y es el tipo de pregunta que una teoría de la conciencia debería poder responder.

Si debemos postular una teoría *ad hoc* de la conciencia para responder esta pregunta, el atractivo de la interpretación de Everett disminuye significativamente. Su mejor característica fue siempre su simplicidad, pero nuevas y arbitrarias leyes psicofísicas la harían tan *ad hoc* como la interpretación de Bohm. Si por otro lado una teoría *independientemente motivada* de la conciencia puede responder esa pregunta, entonces la interpretación de Everett comienza a parecer atractiva.

La teoría de la conciencia que propongo puede responder esa pregunta y ofrecer el tipo correcto de respuesta. La teoría *predice* que un estado cerebral superpuesto debería estar asociado con un número de sujetos distintos de experiencia discreta. Para ver esto, sea un *estado fenoménico maximal* un estado fenoménico que caracteriza toda la experiencia de un sujeto en un momento dado. Sea un *estado físico maximal* un estado físico que caracteriza completamente el estado físico intrínseco de un sistema en un momento dado. Para sacar la conclusión, es suficiente con establecer el siguiente *principio de superposición*:

> Si la teoría predice que un sistema en un estado físico maximal $P$ da origen a un estado fenoménico maximal asociado $E$, entonces la teoría predice que un sistema en una superposición de $P$ con algunos estados físicos ortogonales también dará origen a $E$.

Si este principio es válido, entonces una superposición de estados físicos ortogonales dará origen, cuanto menos, a los estados fenoménicos maximales a los que los estados físicos habrían dado origen separadamente. Esto es precisamente lo que requiere la interpretación de Everett. Si un cerebro está en una superposición de un estado de “percibir arriba” y un estado de “percibir abajo”, entonces dará origen a por lo menos dos sujetos de experiencia, donde uno experimenta un indicador que apunta hacia arriba y el otro experimenta un indicador que apunta hacia abajo. (Por supuesto, estos serán dos sujetos *distintos* de la experiencia, ya que los estados fenoménicos son estados fenoménicos maximales de un sujeto.) Lo mismo es válido para el caso general. Una superposición siempre dará origen al conjunto de sujetos que la interpretación de Everett requiere.

De esta manera, necesitamos establecer que la teoría que formulé implica el principio de superposición. El modo más fácil de ver esto

es apelar al marco teórico del capítulo 9 y, en particular, a la tesis de que la conciencia surge de la implementación de una computación apropiada. Para poder usar esto para establecer el principio, necesitamos determinar que si una computación está implementada por un sistema en el estado físico maximal *P*, también lo está por un sistema en una superposición de *P* con estados físicos ortogonales.

En consecuencia, supóngase que el sistema original (en el estado físico maximal *P*) implementa una computación *C*. Es decir, existe una aplicación entre los subestados físicos del sistema y los subestados formales de *C* tal que las relaciones causales entre los subestados físicos corresponden a relaciones formales entre los subestados formales. Entonces una versión de la misma aplicación también sustentará una implementación de *C* en el sistema superpuesto. Para un subestado determinado *S* del sistema original, podemos encontrar un subestado correspondiente *S'* del sistema superpuesto mediante la obvia relación de proyección: el sistema superpuesto está en *S'* si el sistema obtenido proyectándolo sobre el hiperplano de *P* está en *S*. Debido a que el sistema superpuesto es una superposición de *P* con estados ortogonales, se deduce que si el sistema original está en *S*, el sistema superpuesto está en *S'*. Como la ecuación de Schrödinger es lineal, también se deduce que las relaciones de transición de estados entre los subestados *S'* reflejan precisamente las relaciones entre los subestados originales *S*. Sabemos que estas relaciones a su vez reflejan precisamente las relaciones formales entre los subestados de *C*. Se deduce que el sistema superpuesto también implementa *C*, lo que establece el resultado requerido. Por el principio de invariancia organizacional, si el sistema original da origen a un sujeto de experiencia, el sistema superpuesto dará origen a un sujeto cualitativamente indistinguible de la experiencia.

También podríamos argumentar en favor del principio de superposición aplicando la teoría del doble aspecto de la información y argumentando que la información pertinente incorporada en el estado físico original también está presente en la superposición. Debido a la subdeterminación de esa teoría, sin embargo, este argumento es menos claro que el anterior, de modo que no lo consideraremos aquí. Lo importante es que, de un modo u otro, la teoría de la conciencia que desarrollé parcialmente *predice* el resultado que la interpretación de Everett requiere. Es decir, predice que aunque el mundo está en una superposición gigante, todavía habrá sujetos que experimentan un mundo discreto.

Si no hay otros problemas, se deduce que una combinación de la ecuación de Schrödinger con una teoría independientemente motivada de la conciencia puede predecir nuestra imagen manifiesta del

mundo. Esto es, el único principio físico necesario en la mecánica cuántica es la ecuación de Schrödinger; el postulado de medición y otros principios básicos son equipaje innecesario. Seguramente, también necesitamos principios psicofísicos, pero ellos son necesarios de cualquier forma, y resulta que los principios que son plausibles sobre bases independientes pueden hacer aquí el trabajo requerido. Esto constituye un argumento poderoso en favor de tomar en serio la interpretación de Everett.

## 5. Objeciones a la interpretación de Everett

La interpretación de Everett ha estado sujeta a frecuentes ataques en la literatura; algunas objeciones son más poderosas que otras. Con fines expositivos las agruparé en un cierto número de clases.

### Objeciones basadas en el "desdoblamiento"

Muchas objeciones surgen de interpretar o mal interpretar el enfoque de Everett como un enfoque de "desdoblamiento de mundos". Esto es comprensible, debido a que frecuentemente se la denomina la interpretación de "muchos mundos". Por ejemplo, Bell (1976) objeta que no es claro cuándo debería ocurrir un suceso de "ramificación" debido a la no claridad en la noción de medición, y que no existe ninguna base preferida para la división en mundos. Es evidente que estas objeciones no se aplican a la presente interpretación, que no requiere ninguna "ramificación" objetiva y ninguna base preferida. De modo similar, Hughes (1989) objeta el "chaparrón ontológico" en el proceso de desdoblamiento, y Healey (1984) hace notar que ¡la creación de nuevos mundos viola la conservación de masa-energía! Es una pena que la interpretación de "desdoblamiento" del enfoque de Everett haya ganado tanta aceptación, porque sus dificultades obvias hicieron que la interpretación más interesante no haya recibido la atención que merece.

### Objeciones a una base preferida

Algunas de las objeciones a la interpretación del desdoblamiento de mundos surgen de su necesidad de una base preferida, pero lo mismo ocurre con algunas objeciones de la versión de un único mundo. En particular, surge la pregunta: ¿Por qué las únicas mentes asociadas con un estado cerebral superpuesto corresponden a su descomposición según la base preferida? ¿Por qué no hay mentes que surjan de

otra descomposición o del estado superpuesto como un todo? Esta es una objeción razonable a la propia versión de Everett, la que parece requerir una tal descomposición canónica. Sin embargo, no surge una objeción similar para la versión que formulé, ya que esta implica que una superposición da origen a los sujetos asociados de experiencias discretas sin ninguna necesidad de postular una base preferida. Tampoco tuve ninguna necesidad de recurrir al supuesto de que estas son las *únicas* mentes que el sistema superpuesto origina.

### ¿Qué hay de las mentes superpuestas?

Surge entonces la pregunta, "¿Hay otras mentes asociadas con una superposición?" La respuesta a esto es "tal vez". Si se acepta la teoría del doble aspecto de la información, entonces ya sabemos que puede haber experiencias asociadas con procesos de nivel inferior en un sistema de este tipo. También puede ocurrir que existan sujetos de experiencia asociados con la estructura del procesamiento en una superposición. Tal vez haya mentes asociadas con otras descomposiciones del sistema. Tal vez exista una gran mente superpuesta asociada con el sistema superpuesto completo. La existencia de estas mentes depende de los detalles de una teoría de la conciencia, pero su existencia no parece constituir un problema.

Podríamos tratar de explotar la posibilidad de mentes superpuestas en una objeción a la teoría. Objeción: ¿Por qué *mi* mente no está superpuesta? Respuesta: Porque soy quien soy. La teoría predice que existen mentes no superpuestas y mi mente es una de ellas. Preguntar por qué mi mente no es una de las mentes superpuestas es como preguntar por qué no soy un ratón. Simplemente es parte de la indicatividad primitiva de mi existencia. Las mentes de los ratones existen y las mentes superpuestas podrían existir, pero mi mente no es una de ellas. Objeción: ¿Por qué no tengo acceso a mentes superpuestas, tal como recuerdos de experiencias superpuestas? Respuesta: La teoría predice que las mentes discretas en cuestión experimentarán el mundo en forma totalmente discreta y no tendrán ningún acceso directo a otras partes de la superposición. Todos sus recuerdos serán de observaciones discretas, por ejemplo.

Es discutible, en cualquier caso, que las únicas mentes *interesantes* asociadas con un sistema superpuesto sean el tipo familiar de mentes discretas. Estas mentes son complejas y coherentes; la experiencia refleja la estructura de los procesos racionales. Cualquier otra mente asociada será relativamente incoherente, sin demasiado en el sentido de una estructura interesante. Esta conclusión recibe apoyo del marco de la "descohesión" de Gell-Mann y Hartle (1990) y otros.

Según este marco, la estructura interesante en un sistema adaptativo complejo del tipo ondulatorio se encuentra por lo general dentro de los componentes de una descomposición "natural"; el sistema se "descohesiona" naturalmente a lo largo de ciertas líneas. En sistemas racionales, entonces, la estructura cognitiva coherente puede encontrarse sólo en los componentes de esa descomposición natural, y sólo estos darán origen a mentes coherentes y complejas. Cualquier otro sujeto de experiencia en el sistema no será el tipo de sujeto que califique como persona.

## Objeciones basadas en la identidad personal

Existe un grupo de preocupaciones intuitivas basadas en la identidad del observador. Tómese la mente $M_1$ que recuerdo que estuvo presente a esta hora ayer. Hoy, habrá un gran número de mentes que descienden de ella, en diferentes "ramas" de la superposición. Mi mente $M_2$ es sólo una de ellas. Podría entonces preguntar: ¿Por qué terminé *aquí*, en lugar de en una de las otras ramas? Como Hofstadter (1985b) señala:

> ¿Por qué mi sensación unitaria de mí mismo se propaga hacia abajo a lo largo de *esta* rama aleatoria en lugar de a lo largo de alguna otra? ¿Qué *ley* subyace en las elecciones aleatorias que seleccionan la rama que siento que sigo? ¿Por qué mi sensación de mí mismo no acompaña a los otros yo cuando se dividen, siguiendo otras rutas? ¿Qué asocia *yoidad* al punto de vista de este cuerpo que evoluciona a lo largo de esta rama del universo en este momento temporal?

A esto, debemos nuevamente invocar la indicatividad primitiva: mi mente es *esta* y eso es todo. Existe una sensación de que algo más profundo debe estar ocurriendo y que de algún modo es un hecho profundo acerca del mundo que la mente $M_1$ de ayer evolucionó en la mente $M_2$ de hoy en día y no en una de las otras. Pero desde un punto de vista objetivo, no hay nada especialmente privilegiado en esta rama. Aun desde el punto de vista de $M_1$, todas las mentes de hoy son igualmente privilegiadas. Ninguna de ellas es la única heredera legítima de $M_1$; todas ellas portan la "yoidad" de $M_1$ en el mismo grado. Es sólo desde *este* punto de vista, el punto de vista de $M_2$, que $M_2$ parece privilegiado (por supuesto, mis contrapartes en otros lados de la superposición tienen la misma sensación acerca de ellos mismos). Este papel privilegiado de $M_2$ es sólo otro fenómeno indicativo, como el hecho de que yo soy David Chalmers y no Rolf Harris. *Esta* mente está aquí y no allí. Esto es tan problemático como cualquier

hecho indicativo, pero no existe ninguna asimetría ulterior en el mundo.

Existe una fuerte intuición de que siempre debe haber un hecho empírico acerca de la identidad personal: si existen numerosas mentes que descienden de mi estado actual, debe existir un hecho acerca del cual una de ellas seré *yo*. Pero esta idea ha estado sometida a una poderosa crítica por Parfit (1984), quien argumenta persuasivamente que no hay nada más en el hecho de la identidad personal que hechos como los de continuidad psicológica, memoria, y similares. Si aceptamos este análisis, entonces cada una de las mentes de mañana será una candidata equivalente para ser *yo*, y no hay ningún hecho que permita distinguirlas. Hay algo perturbador en esta conclusión, que reduce el "flujo" determinado de la identidad personal a una ilusión, pero el análisis de Parfit ofrece razones para creer que este flujo determinado fue siempre una ilusión.

## La interpretación de las probabilidades

La objeción más sustancial a la interpretación de Everett es que no puede interpretar las *probabilidades* que produce el postulado de medición.[8] En un caso determinado el postulado de medición podría decirnos que al hacer una cierta medición, habrá una probabilidad de 0,9 de encontrar un indicador "arriba" y una probabilidad de 0,1 de encontrar un indicador "abajo". Según la interpretación de Everett, lo que realmente ocurre es que el indicador y el estado cerebral de un observador entran en una superposición, lo que resulta en (al menos) dos sujetos de experiencia. Uno de estos tiene una experiencia de un indicador "arriba" y otro experimenta un indicador que apunta hacia abajo. Exactamente lo mismo habría ocurrido si las probabilidades hubieran sido 50:50. Es verdad que en el caso 90:10, la mayor parte de la *amplitud* de la función de onda superpuesta se concentra en el área del estado cerebral "arriba", pero ¿qué tiene esto que ver con las probabilidades?

Everett enfrenta esta pregunta mediante la formulación de una *medida* sobre el espacio de observadores que corresponde a las probabilidades producidas por el postulado de medición (esto es, corresponde al cuadrado de la amplitud de la parte correspondiente de la función de onda). Utilizando esta medida, argumenta que, en el límite, la *mayor parte* de los observadores (esto es, un subconjunto de observadores con medida uno) tendrán recuerdos de observaciones que concuerdan con las frecuencias predichas por las probabilidades en el postulado de medición. Por ejemplo, entre observadores que hicieron una medición como la que describimos más arriba muchas

veces, la mayor parte de ellos recordará haber encontrado un puntero "arriba" el 90% del tiempo y un puntero "abajo" el 10% del tiempo. De esta forma se le asigna un papel a las probabilidades. Sin embargo, surge la pregunta: ¿Qué es lo que justifica estas medida sobre el espacio de observadores? Si midiésemos el espacio de modo diferente, entonces podrían surgir frecuencias muy diferentes. Por ejemplo, si asignásemos medidas iguales cada vez que surgen dos observadores de una superposición (sin importar la amplitud), entonces la mayoría de los observadores recordaría una tasa de "arriba"- "abajo" de 50:50. Ni la ecuación de Schrödinger, ni las leyes psicofísicas sostienen que alguna de estas medidas sea la "correcta".

Albert y Loewer (1988) responden a esta preocupación prescindiendo de las medidas. En cambio, postulan leyes psicofísicas más radicales según las cuales existe una infinidad de mentes asociadas con cada estado cerebral. Para cada mente postulada por el enfoque previo, esta teoría postula un conjunto infinito de mentes cualitativamente idénticas. Más aún, dondequiera que la teoría de Everett prediga que una mente divergirá en dos mentes, esta teoría dice que cualquier mente dada irá en una dirección o en la otra, con probabilidades determinadas por el postulado de medición. Así, si tomamos una mente arbitraria asociada con el estado cerebral antes de la medición de más arriba, tendrá una probabilidad del 90% de evolucionar en un estado "percibiendo arriba" y una probabilidad de un 10% de evolucionar en un estado "percibiendo abajo". De este modo se preservan las predicciones probabilísticas del cálculo de la mecánica cuántica.

Claramente hay aquí una pérdida de simplicidad. Las nuevas leyes psicofísicas no tienen ninguna motivación independiente y la teoría también necesita leyes "intrapsíquicas" extra que gobiernen la *evolución* de las mentes. Al adoptar estos postulados *ad hoc*, la teoría sacrifica algunas de las virtudes fundamentales de la interpretación de Everett. También puede sostenerse que las leyes intrapsíquicas son problemáticas, en el sentido de que postulan hechos irreducibles profundos acerca de la identidad personal en el tiempo. Es difícil saber qué hacer con estos hechos. Aceptarlos requeriría desechar el análisis de Parfit de la identidad personal, por ejemplo. No supervienen ni siquiera naturalmente a los hechos físicos, de modo que complican la imagen metafísica. Esta interpretación debe tenerse en mente como una posibilidad, pero posee un costo significativo.

La alternativa es prescindir del aparato extra y ver si las probabilidades pueden recuperarse de algún otro modo. Es tentador ver esto como un problema acerca de la indicatividad. ¿Por qué de

todos los lugares en la función de onda en los que podría haber terminado, lo hice en una región en la que mis recuerdos concuerdan con las predicciones del cálculo? Una posibilidad es simplemente tomar esto como un hecho indicativo primitivo: *algunas* mentes están en esta área, y ocurre que yo soy una de ellas. Pero esto no parece satisfactorio ya que la notable regularidad del cálculo resulta entonces ser una enorme suerte. Lo que necesitamos es algún modo de argumentar que no existe esa suerte.

Aun al notar que es una suerte que yo haya terminado *aquí*, está implícita la idea de que existe algún tipo de medida sobre el espacio de mentes. La sugerencia es que es antecedentemente más probable que termine siendo una mente de un tipo en lugar de otro, quizá debido a la abundancia relativa de esas clases. Este tipo de medida implícita está presente en gran parte de nuestro razonamiento acerca del mundo. Cuando razono inductivamente desde alguna evidencia a una conclusión, yo sé que para *algunos* observadores en una posición epistémica similar la conclusión no será válida, pero supongo que para la *mayor parte* de estos observadores la conclusión será válida, aunque exista un número infinito en cada clase. Es decir, supongo que es antecedentemente más probable que yo esté en una clase y no en otra. Este tipo de razonamiento supone implícitamente alguna clase de medida sobre el espacio de las mentes.

Tal vez, entonces, podamos justificar las probabilidades introduciendo explícitamente este tipo de medida. El grueso de la amplitud de la función de onda está concentrado en áreas en las que los recuerdos de los observadores concuerdan con las predicciones del cálculo. Tal vez sea más probable que mi mente resulte estar en un área de alta amplitud que en un área de baja amplitud. En particular, si suponemos que la probabilidad antecedente de que yo resulte ser una mente en lugar de otra es proporcional al cuadrado de la amplitud de la parte asociada de la función de onda, entonces se deduce que casi seguramente tendré recuerdos en las frecuencias predichas por el cálculo de la mecánica cuántica.

Pero, ¿a qué corresponde objetivamente esta medida? ¿Debe considerarse un hecho básico acerca de la distribución de los yo? ¿Puede de algún modo justificarse como la medida canónica en este espacio? Estas son preguntas difíciles que están estrechamente ligadas al misterio de la propia indicatividad: ¿por qué resulté yo ser *esta* persona en lugar de alguna otra? Este es uno de los misterios básicos, y es muy poco claro cómo exactamente debería dársele una respuesta. Sin embargo, la idea de una medida sobre el espacio de las mentes parece prometedora e incluso podría ser necesaria para algunos otros propósitos, tal como la justificación de la inducción.

Mientras tanto, la interpretación de las probabilidades sigue siendo la dificultad más significativa para la interpretación de Everett.

## 6. Conclusiones

Debe admitirse que la interpretación de Everett es casi imposible de creer. Postula que existe muchísimo más en el mundo que aquello de lo que nos percatamos. Según esta interpretación, el mundo está realmente en una superposición gigante de estados que han evolucionado de diferentes maneras desde el comienzo del tiempo, y sólo experimentamos el subestado más pequeño del mundo. También postula que mi futuro no está determinado: dentro de un minuto, habrá un número grande de mentes con el mismo derecho a ser *yo*. Un minuto pasó desde que escribí la última oración; ¿quién puede saber qué están haciendo ahora todas esas otras mentes?

Por otro lado, es claro en este momento que *todas* las interpretaciones de la mecánica cuántica son, en alguna medida, descabelladas. Esta es la paradoja fundamental de la mecánica cuántica. Los tres principales candidatos para interpretación son quizá la interpretación de Wigner según la cual la conciencia provoca el colapso, la interpretación de variables ocultas no locales de Bohm y la interpretación de Everett. De estas, la interpretación de Wigner implica que los objetos macroscópicos están frecuentemente en superposiciones hasta que una mirada casual de un observador provoca que colapse. El enfoque de Bohm implica que la trayectoria de cada partícula en el universo depende del estado de todas las demás. Y el enfoque de Everett implica que hay mucho más en el mundo de lo que alguna vez supusimos.

De estas, quizás el enfoque de Bohm sea el menos descabellado, el de Everett el más descabellado, y el de Wigner estaría en el medio. Ordenados según su mérito teórico, por otro lado, se invierte la secuencia. El enfoque de Bohm es insatisfactorio debido a su naturaleza compleja y "arreglada". El enfoque de Wigner es bastante elegante, con sus dos leyes dinámicas básicas que reflejan el cálculo de la mecánica cuántica, si pueden resolverse todos los detalles. Pero el enfoque de Everett es, de lejos, el más simple. Sólo postula la ecuación de Schrödinger, el principio que todas las interpretaciones de la mecánica cuántica aceptan. También, tiene las virtudes de ser una teoría totalmente local y directamente compatible con la teoría de la relatividad, virtudes que no encontramos en las otras interpretaciones.

También vale la pena notar que las otras dos interpretaciones contienen elementos de lo que resulta contrario a la intuición en la

interpretación de Everett. Según el enfoque de Wigner, debemos aceptar que el universo evolucionó en una superposición gigante al estilo de Everett —tal vez con estrellas superpuestas y rocas superpuestas, si no con gatos superpuestos—, al menos hasta que la primera entidad consciente evolucionó para colapsar la función de onda. Según el enfoque de Bohm, la función de onda no colapsada de Everett sigue presente como la "onda piloto" que guía la posición de las diversas partículas. Toda la estructura que está presente en otros componentes sigue entonces presente en el estado del mundo, aun cuando la mayor parte sea irrelevante para la evolución de las partículas. Dado que estos enfoques también requieren una función de onda no colapsada en papeles centrales, se podría argumentar que la relativa inverosimilitud del enfoque de Everett disminuye.

Por supuesto, siempre es posible que pueda desarrollarse una nueva teoría que supere a todas estas en plausibilidad y méritos teóricos. Pero no parece particularmente probable. La completa ausencia de anomalías experimentales sugiere que el cálculo de la mecánica cuántica está aquí para quedarse como teoría predictiva. Si esto es así, no podemos esperar que nuevos desarrollos empíricos resuelvan el problema. Tal vez sean los desarrollos conceptuales los que lleven a una interpretación nueva y mejorada, pero podría ocurrir que en este momento los nichos más prometedores en el espacio conceptual ya hayan sido explotados. De ser así, podríamos encontrarnos fijados al espectro actual de opciones; tal vez podamos hacer refinamientos significativos, pero las ventajas y desventajas serían de un tipo cualitativamente similar. De estas opciones, la interpretación de Everett parece ser, en muchas formas, la más atractiva, pero al mismo tiempo la más difícil de aceptar.

A lo largo de esta obra propuse algunos enfoques contrarios a la intuición. Durante mucho tiempo me resistí al dualismo mente-cuerpo, pero llegué a un punto en el que terminé aceptándolo, no sólo como el único enfoque defendible sino como un enfoque satisfactorio por derecho propio. Siempre es posible que me haya confundido, o que exista alguna nueva y radical posibilidad que pude haber pasado por alto; pero puedo decir tranquilo que creo que es muy probable que el dualismo sea verdadero. También planteé la posibilidad de una especie de panpsiquismo. Al igual que el dualismo mente-cuerpo, este es inicialmente contrario a la intuición, pero esta sensación desaparece con el tiempo. No estoy seguro de si el enfoque es verdadero o falso, pero al menos es intelectualmente atractivo y, si se reflexiona sobre él, no demasiado descabellado como para que no pueda ser aceptable.

La extravagancia de la interpretación de Everett es de un orden de magnitud diferente. La encuentro la interpretación más intelec-

tualmente atractiva entre las interpretaciones de la mecánica cuántica, pero confieso que no puedo creer en ella incondicionalmente. Si Dios me forzase a apostar la vida a la verdad o falsedad de las doctrinas que propuse, apostaría bastante confiado que la experiencia es fundamental y débilmente a que la experiencia es ubicua. Pero, en lo que respecta a la interpretación de Everett estaría indeciso y quizá no tendría el valor suficiente para finalmente apostar por ella.[9] Tal vez sólo sea demasiado extraña para creer. No obstante, no es claro en el análisis final si debería darse mucho peso a estas dudas intuitivas. El enfoque es simple y elegante, y predice que habrá observadores que ven el mundo exactamente como yo lo veo. ¿No es eso suficiente? Podría ocurrir que nunca podamos aceptar emocionalmente el enfoque, pero al menos deberíamos tomar en serio la posibilidad de que sea verdadero.

# Notas

## Capítulo 1

1. Véase Nagel, 1974. La primera utilización de esta frase en contextos filosóficos se atribuye usualmente a Farrell (1950). Véase también Sprigge, 1971.

2. Distintos autores utilizan el término "qualia" de diferentes modos. Yo utilizo el término en lo que creo es el modo estándar, para referir a aquellas propiedades de los estados mentales que tipifican esos estados según cómo sea experimentarlos. Al utilizar el término, no deseo comprometerme inmediatamente con cuestiones ulteriores, tales como si los qualia son incorregiblemente cognoscibles, si son propiedades intencionales, etcétera. Los qualia pueden ser propiedades de estados mentales "internos" además de sensaciones. Suele ser conveniente hablar como si los qualia fuesen propiedades instanciadas directamente por un sujeto, en lugar de por los estados mentales de ese sujeto; esta práctica es inocua y está justificada por el hecho de que los qualia corresponden a tipos de estados mentales por derecho propio.

3. Utilizo expresiones como "sensación de rojo", "sensación de verde", y similares a través del libro. Por supuesto al hacerlo no deseo sugerir que las experiencias instancian el mismo tipo de propiedades del color que son instanciadas por objetos (manzanas, árboles) en el mundo externo. Esta forma de expresión siempre puede reformularse como "experiencia del tipo que usualmente tengo (en el mundo real) cuando miro objetos rojos", etcétera, pero la locución más breve resulta más natural.

4. Cocine 12 tazas de frijoles de carete en agua hirviendo a la que se le agregaron previamente cuatro cucharadas soperas de sal. Cocine hasta que los frijoles estén tiernos y luego sumerja en agua fría. Mezcle dos pimientos rojos cortados en trozos, 5 pimientos verdes cortados en trozos, 2 cebollas grandes cortadas en trozos, 3 tazas de pasas de uva y un manojo de cilantro picado con un aderezo hecho con $1\frac{1}{2}$ taza de aceite de maíz, $\frac{3}{4}$ de taza de vinagre blanco, 4 cucharadas soperas de azúcar, y 1 cucharada sopera de sal, 4 cucharadas de pimienta negra, 5 cucharadas soperas de curry en polvo y

media cucharada sopera de clavo de olor molido. Sirva frío. Agradezco a Lisa Thomas y al Encore Café.

5. Para un número abundante de reflexiones sobre variedades específicas de experiencias, véase *A Natural History of Senses* de Ackerman (1990). Este texto proporciona material para que aquellos que están absorbidos por su experiencia consciente puedan reflexionar durante días.

6. Es interesante que Descartes solía excluir a las sensaciones de la categoría de lo mental y las asimilaba, en cambio, a lo corporal; de este modo, no todo estado fenoménico (al menos, tal como yo entiendo la noción) sería considerado mental.

7. Esta interpretación común de Ryle no hace justicia a la sutileza de sus enfoques pero es, al menos, una ficción útil.

8. Existen otras formas de funcionalismo, como la desarrollada por Putnam (1960). No las considero aquí, ya que fueron formuladas como hipótesis empíricas más que como análisis de los conceptos mentales.

9. Nagel (1970) señala un punto similar en contra de Armstrong, con referencia al problema de otras mentes.

10. El argumento de Searle depende de la tesis de que sin la conciencia, no hay ningún modo de explicar la "forma aspectual" que la intencionalidad exhibe, como cuando alguien cree algo acerca de Venus bajo su aspecto de "estrella matutina" pero no bajo su aspecto de "estrella vespertina". No me resulta claro que la forma aspectual no pueda explicarse de otras formas; se podría incluso argumentar que los ordenadores actuales exhiben algo similar, cuando almacenan información acerca de mí bajo un "aspecto" pero no otro (por ejemplo, bajo mi nombre pero no bajo mi número de seguridad social). Podría objetarse que esta es sólo una forma aspectual "como si", ya que la única verdadera forma aspectual es la forma aspectual *fenoménica*; pero esto parecería trivializar el argumento.

11. Algunas observaciones de Lockwood (1989) y Nagel (1986) sugieren una posición semejante, aunque no estoy seguro de que alguno de los dos acepte esa posición.

12. Algunos podrían señalar argumentos como el de Kripke (1982) al efecto de que el contenido de una creencia no está determinado por propiedades psicológicas o fenoménicas. Estos argumentos son discutibles pero, en cualquier caso, es notable que su consecuencia no es que el contenido sea un nuevo elemento irreducible de la mente, sino que el contenido mismo es indeterminado. Lo que ocurre aquí es que consideraciones como las del texto nos dan buenas razones para creer que no existe ningún tercer aspecto de la mente que pueda variarse en forma independiente; de modo que cualquier cuestión que no pudiese decidirse mediante los primeros dos elementos no podría decidirse.

13. Este es un análisis de "tema neutro" de las nociones fenoménicas específicas, no demasiado diferente de los propuestos por Place (1956) y Smart (1959). Algo es una experiencia de naranja, en forma muy aproximada, si es el tipo de experiencia causada por lo general por las naranjas. Place: "Cuando describimos la postimagen como verde .... decimos que tenemos

el tipo de experiencia que normalmente tenemos cuando miramos un fragmento verde de luz y que hemos aprendido a describir de esa manera", (p. 49); Smart: "Cuando una persona dice 'Veo una postimagen naranja amarillenta', está diciendo algo así como: '*Ocurre algo que es como lo que ocurre cuando* tengo mis ojos abiertos, estoy despierto, y hay una naranja iluminada bajo una buena luz frente a mí'" (p. 150). Pero debido a la ocurrencia de la noción no analizada de "experiencia", este análisis no es suficiente para establecer inmediatamente una identificación entre los estados fenoménicos y los físicos, al modo como Place y Smart lo sugirieron. La concepción de Smart evita este problema dejando a la "experiencia" fuera del análisis al recurrir a la frase equívoca "ocurre algo". Si "ocurre algo" se interpreta en una forma lo suficientemente amplia como para cubrir cualquier tipo de estado, entonces el análisis es inadecuado; si se lo interpreta de un modo restringido como un tipo de experiencia, el análisis está más cerca del objetivo pero no basta para la conclusión.

14. Jackendoff distingue la "mente fenomenológica" de la "mente computacional". Esta distinción es muy similar a la distinción fenoménica-psicológica esbozada aquí, aunque no me gustaría dar por sentado que los procesos psicológicos son computacionales.

15. Nelkin (1989) distingue CN (conciencia en el sentido de "Nagel") de C1 (un estado de procesamiento de información de primer orden) y C2 (acceso no inferencial directo de segundo orden a otros estados conscientes). En otro artículo (Nelkin, 1993), hace una distinción relacionada entre fenomenalidad, intencionalidad e introspectibilidad. Bisiach (1988) distingue C1 (experiencia fenoménica) de C2 (el acceso de partes o procesos de un sistema a otras de sus partes o procesos). Natsoulas (1978) distingue un gran número de sentidos del término "conciencia". Dennett (1969) distingue dos clases de "percatación", la primera asociada con los informes verbales y la segunda más generalmente con el control de la conducta, aunque ninguna de estas es claramente una noción fenoménica.

16. Rosenthal separa explícitamente la conciencia de la "cualidad sensorial", y dice que formula una teoría sólo de la primera, lo que podría sugerir que los aspectos fenoménicos no están en discusión. Pero también dice que un estado es consciente cuando existe algo que es como estar en ese estado, lo que sugiere que el sujeto es conciencia fenoménica después de todo. Sin embargo, hay muy poco en la concepción de Rosenthal que sugiera una *explicación* de la conciencia fenoménica. ¿Por qué debería la existencia de un pensamiento de orden superior acerca de un estado llevar a que haya algo que es como estar en ese estado? Aparte de argumentar que es plausible que los dos fenómenos vayan juntos en la práctica, Rosenthal no da ninguna respuesta a esta cuestión.

17. Una excepción es el campo de la psicofísica; es posible que arroje luz sobre las diversas características de la experiencia consciente, aunque no proporcionará una explicación completa. Analizaré esto más en profundidad en el capítulo 6.

## Capítulo 2

1. La idea de superveniencia fue introducida por Moore (1922). El término fue publicado por primera vez en Hare (1952). Davidson (1970) fue el primero en aplicar la noción al problema mente-cuerpo. Más recientemente, Kim (1978, 1984, 1993), Horgan (1982, 1984c, 1993), Hellman y Thompson (1975), y otros desarrollaron una teoría sofisticada de la superveniencia.

2. Utilizo "hecho A" como abreviatura de "instanciación de una propiedad A". La apelación a los hechos hace que la discusión sea menos difícil, pero todo discurso sobre hechos y sus relaciones puede finalmente traducirse en términos de patrones de coinstanciación de propiedades; daré los detalles en notas, cuando ello sea necesario. En particular, debe notarse que la identidad del individuo que instancia una propiedad A es irrelevante para un hecho A tal como lo interpreto; todo lo que importa es la instanciación de la propiedad. Si la identidad de un individuo fuese parcialmente constitutiva de un hecho A, entonces cualquier hecho A implicaría hechos acerca de las propiedades esenciales de ese individuo, en cuyo caso la definición de superveniencia llevaría a consecuencias contrarias a la intuición.

3. Supongo, tal vez artificialmente, que los individuos tienen fronteras espaciotemporales precisas, de modo que sus propiedades físicas consisten en las propiedades instanciadas en esa región del espacio-tiempo. Si debemos considerar que los objetos espacialmente distintos son físicamente idénticos para los propósitos de la superveniencia local, entonces cualquier propiedad que involucre a la posición espaciotemporal absoluta debe omitirse de la base de superveniencia (aunque podríamos evitar la necesidad de recurrir a objetos espacialmente distintos si sólo consideramos objetos *posibles* con la misma posición). También, siempre hablo como si el mismo tipo de individuo instanciase propiedades de bajo nivel y de alto nivel, de modo que una mesa, por ejemplo, instancia propiedades microfísicas en virtud de estar caracterizada por una distribución de dichas propiedades. Tal vez fuese más estrictamente correcto hablar de que las propiedades microfísicas son instanciadas sólo por entidades microfísicas, pero mi forma de expresión simplifica las cosas. De cualquier forma, las cuestiones verdaderamente centrales involucrarán la superveniencia global más que la local.

4. Existen diversos modos de especificar precisamente qué significa que dos mundos sean idénticos respecto de un conjunto de propiedades; esto no importará mucho para el análisis. Quizá lo mejor sea decir que dos mundos son idénticos respecto de sus propiedades A si existe una aplicación biunívoca entre las clases de individuos que instancian propiedades A en ambos mundos, tal que dos individuos correspondientes cualesquiera instancian las mismas propiedades A. Para los propósitos de la superveniencia global necesitamos entonces estipular que las aplicaciones mediante las cuales dos mundos se consideran A y B indiscernibles son compatibles entre sí; es decir, ningún individuo se vincula con su contraparte bajo la aplicación A pero con otro bajo la aplicación B. La definición de superveniencia global adopta la siguiente forma. Dos mundos cualesquiera que son A idénticos (bajo una aplicación) son B idénticos (bajo una extensión de esa aplicación).

Un modo más común de hacer esto es estipular que los mundos A idénticos deben contener exactamente los mismos individuos que instancian las mismas propiedades, pero como McLaughlin (1995) señala, esto es irrazonablemente fuerte: asegura que cosas como la cardinalidad del mundo y las propiedades esenciales de los individuos supervengan a cualquier propiedad. La definición que propongo evita este problema, ya que hace que *sólo* patrones de propiedades A y ninguna otra cosa participen de la relación de determinación.

5. Con una excepción: Dios no podría haber creado un mundo que no fuese creado por Dios, ¡aun cuando podemos suponer que un mundo no creado por Dios es lógicamente posible! Ignoraré este tipo de complicación.

6. La relación de este tipo de posibilidad con la deducibilidad en los sistemas formales es sutil. Puede sostenerse que los axiomas y las reglas de inferencia de sistemas formales específicos se justifican precisamente en términos de una noción previa de posibilidad y necesidad lógica.

7. La noción intuitiva de posibilidad natural es conceptualmente previa a la definición en términos de leyes naturales: puede considerarse que una regularidad es una ley sólo en el caso de que es válida en todas las situaciones que puedan surgir en la naturaleza; es decir, en todas la situaciones que son naturalmente posibles en el sentido intuitivo. Tal como a veces se lo formula, para que algo pueda ser considerado una ley debe ser válido no sólo en las situaciones reales sino también en las contrafácticas; se requiere la noción más básica de posibilidad natural para determinar qué situaciones contrafácticas son pertinentes.

8. Los términos "necesidad física" y "necesidad causal" suelen utilizarse también para seleccionar aproximadamente este tipo de necesidad, pero no deseo presuponer que todas las leyes de la naturaleza son físicas o causales.

9. La importante distinción entre la superveniencia lógica y natural suele ocultarse o ignorarse en la literatura, donde es frecuente que no se especifique la modalidad de las relaciones de superveniencia. La superveniencia natural (o nomológica) sin superveniencia lógica es analizada en Van Cleve (1990), quien la utiliza para explicitar una variedad de emergencia. Seager (1991) expone una distinción relacionada entre lo que llama superveniencia *constitutiva* y *correlativa*. Estas corresponden de un modo directo a la superveniencia lógica y natural, aunque Seager no analiza las nociones del mismo modo.

10. La superveniencia débil sólo requiere que "ninguna diferencia B sin una diferencia A" sea válida dentro de un mundo, no a través de mundos (véase Kim [1984] para los detalles). La falta de fuerza modal en esta relación la hace demasiado débil para la mayoría de los propósitos. En el mejor de los casos, podría tener un papel en la expresión de las restricciones conceptuales sobre el discurso no fáctico (como en Hare, 1984), aunque como Horgan (1993) señala, aun estas restricciones parecen involucrar la dependencia a través de mundos. Seager (1988) apela a la superveniencia débil para expresar un tipo de correlación sistemática intramundo que no es estrictamente necesaria, pero la superveniencia natural sirve a este propósito mucho mejor.

11. La superveniencia natural global sin regularidad localizada es una

noción coherente en una concepción no humeana de las leyes, aunque quizá no en una concepción humeana (basada en regularidades). Aun en una concepción no humeana, sin embargo, es difícil ver en qué podría consistir la evidencia en favor de este tipo de relación.

12. Horgan (1982), Jackson (1994) y Lewis (1983b) estudian un problema relacionado en el contexto de definir el materialismo.

13. La definición revisada puede detallarse más precisamente según las líneas de la nota 4. Sea $B(W)$ la clase de individuos con propiedades $B$ en un mundo $W$. Podemos decir que $W'$ es *B-superior* a $W$ si existe una aplicación inyectiva $f$: $B(W) \rightarrow B(W')$ (esto es, una aplicación biunívoca de $B(W)$ en un subconjunto de $B(W')$) tal que para todo $a \in B(W)$, $f(a)$ instancia toda propiedad $B$ que $a$ instancia. Entonces, las propiedades $B$ supervienen lógicamente a las propiedades $A$ en $W$ si todo mundo que es $A$-indiscernible de W es $B$-superior a $W$, donde las aplicaciones $B$ pertinentes están nuevamente constreñidas a ser extensiones de las aplicaciones $A$.

Para ver que la restricción es necesaria, imagínese que nuestro mundo tenga un número infinito numerable de mentes psicológicamente idénticas, de las cuales una está realizada en ectoplasma y el resto están físicamente realizadas. Intuitivamente, lo psicológico no superviene a lo físico en este mundo, pero todo mundo físicamente indiscernible es psicológicamente superior. Aunque esperamos que el mundo libre de ectoplasma contradiga la superveniencia, existe una aplicación biunívoca entre las psicologías en ese mundo y en nuestro mundo. El problema es que esta aplicación no respeta la correspondencia física, ya que aplica una entidad física en una entidad ectoplasmática; por eso necesitamos la nueva restricción.

14. Para los propósitos de esta definición, la relación de contención entre mundos puede considerarse primitiva. Lewis (1983a) y Jackson (1993) notaron que es infructuoso analizar este tipo de noción para siempre. Algo debe considerarse primitivo, y la relación de contención parece tan clara como cualquiera. Algunos podrían preferir hablar no de mundos que contienen a $W$ como una parte propia, sino de mundos que contienen un duplicado cualitativo de $W$ como una parte propia; esto funciona igualmente bien.

15. Nótese que según esta definición, existen hechos positivos que no son instanciaciones de propiedades positivas. Piénsese en las instanciaciones de la propiedad "no tener hijos o ser un canguro", por ejemplo. Tal vez los hechos positivos deberían definirse más estrictamente como instanciaciones de propiedades positivas, pero hasta donde puedo decirlo la definición más débil no tiene efectos perjudiciales.

16. Supuestamente, la superveniencia lógica de propiedades en nuestro mundo debería ser una tesis *legaliforme*. Si fuese el caso de que *hubiera* habido ángeles vivientes no físicos si las cosas se hubiesen desarrollado en forma un poco diferente en nuestro mundo (quizás unas pocas fluctuaciones aleatorias diferentes), aun cuando las leyes de la naturaleza se hubiesen respetado, entonces habría sido un mero accidente de la historia que las propiedades biológicas fueran lógicamente supervenientes a las propiedades físicas. Se obtiene una tesis metafísica más fuerte e interesante si reemplazamos las referencias a nuestro mundo y a los individuos reales en las

definiciones de la superveniencia lógica por una referencia a mundos e individuos naturalmente posibles. Esto permite desechar escenarios como el de más arriba. Como bonificación, permite además decidir si las propiedades no instanciadas, como la de ser un rascacielos de dos kilómetros de alto, son lógicamente supervenientes o no. Según la definición previa, estas propiedades supervendrían en forma vacía.

Esto produce la siguiente definición: Las propiedades *B* son lógicamente supervenientes a las propiedades *A* si para cualquier situación naturalmente posible *X* y cualquier situación lógicamente posible *Y*, si *X* e *Y* son *A*-indiscernibles entonces *Y* es *B*-superior a *X* (con la restricción usual). O más brevemente: para cualquier situación naturalmente posible, los hechos *B* acerca de esa situación están implicados por los hechos *A*. Esta modificación no hace ninguna diferencia significativa para el análisis en el texto, de modo que la omitiré en beneficio de la simplicidad. La exposición puede fácilmente reformularse en términos de la definición más estricta sólo reemplazando las referencias pertinentes a "nuestro mundo" por referencias a "todos los mundos naturalmente posibles" a través del texto sin, por lo general, ninguna pérdida en la plausibilidad de las aseveraciones asociadas.

El resultado se parece a la definición estándar de superveniencia local "fuerte" (Kim, 1984), en la cual hay dos operadores modales. Según esa definición, las propiedades *B* supervienen a las propiedades *A* si necesariamente, para todo *x* y toda propiedad-*B F*, si *x* posee *F*, entonces existe una propiedad-*A G* tal que *x* posee *G*, y necesariamente si cualquier *y* posee *G* también posee *F*. (Si es necesario, las propiedades-A *G* pueden pensarse como complejos de propiedades *A* más simples.) La cuestión de los ángeles deja en claro que el primer operador modal debería siempre interpretarse como necesidad natural, aun cuando el segundo sea de necesidad lógica. La definición estándar de superveniencia global (que los mundos A-indiscernibles son B-indiscernibles) no resulta tan bien, y debe ser modificada en la forma que sugerí. De ser necesario, puede formularse una definición paralela de superveniencia "metafísica". Por supuesto, los problemas con los ángeles no surgen en el caso de la superveniencia natural, ya que no hay ninguna razón para creer que el ectoplasma sea naturalmente posible, de modo que la definición directa de superveniencia natural es satisfactoria.

17. Posiblemente, deberíamos utilizar la definición más fuerte de superveniencia lógica, de modo que el materialismo es verdad si todos los hechos positivos acerca de los mundos *naturalmente posibles* están implicados por los hechos físicos acerca de esos mundos. Tómese un mundo libre de ectoplasma en el cual, sin embargo, el ectoplasma no físico es una posibilidad natural; tal vez hubiese podido evolucionar si algunas pocas fluctuaciones aleatorias hubieran sido diferentes. Parece razonable decir que el materialismo es falso en un mundo así o, al menos, que es verdad sólo en un sentido débil.

18. Para obtener la equivalencia, necesitamos el principio plausible de que si un mundo *A* es una parte propia de un mundo *B*, entonces existe algún hecho positivo que vale en *B* pero que no vale en *A*; es decir, existe algún hecho que vale en *B* y en todos los mundos mayores pero que no vale en *A*.

19. También se reduce a algo similar a las definiciones de Horgan (1982) y Lewis (1983b) pero, a diferencia de estas, no se basa en la noción algo oscura de "propiedad foránea" para descartar los mundos ectoplasmáticos del dominio de mundos posibles pertinentes.

20. En la literatura filosófica suele señalarse la realizabilidad múltiple como el principal obstáculo para la "reducción" pero, como Brooks (1994) argumenta, esto parece relativamente irrelevante para el modo como se utilizan las explicaciones reductivas en las ciencias. Fenómenos biológicos como las alas pueden realizarse de muchos modos diferentes, por ejemplo, pero los biólogos, de todas formas, formulan explicaciones reductivas. Como lo señalaron Wilson (1985) y Churchland (1986), muchos fenómenos físicos que suelen considerarse paradigmas de reducibilidad (por ejemplo, la temperatura) son, de hecho, múltiplemente realizables.

21. Algunos dirían que no deberíamos hablar de a qué podría referirse "agua" si el mundo XYZ resultase real porque, en ese caso, ¡la palabra que suena como "agua" sería una palabra totalmente diferente! Si esto nos preocupa, podemos simplemente hablar de a qué se referiría la palabra homófona; o mejor, podríamos considerar estos escenarios como posibilidades *epistémicas* (en un sentido amplio) y los condicionales como condicionales epistémicos, para evitar las preocupaciones acerca de las propiedades esenciales de las palabras. En cualquier caso, la cuestión general de que la referencia en el mundo real depende de cómo este resulta es clara independientemente de cómo describamos el escenario. En el texto, en general ignoraré esta sutileza.

22. No todos están convencidos de que Kripke y Putnam estén en lo correcto en su afirmación de que el agua es necesariamente H2O y que XYZ no es agua; para algunas dudas, véase Lewis, 1994. Parecería que la *mayoría* de nuestras prácticas lingüísticas serían exactamente las mismas aun si usásemos el término "agua" para seleccionar la sustancia acuosa en mundos contrafácticos. El hecho de que es fácil inclinarse en cualquiera de los dos sentidos respecto de esta cuestión sugiere que nada realmente importante para la explicación de un fenómeno como el agua gira en torno de la naturaleza de la intensión secundaria. El hecho de que siempre podemos utilizar términos como "sustancia acuosa" en lugar de "agua" en estas cuestiones es un indicio de que es improbable que la necesidad *a posteriori* pueda cambiar algo que sea realmente importante en temas como los de la explicación, el fisicalismo y similares. Siewert (1994) evita de este modo las cuestiones acerca de la necesidad *a posteriori*. Yo también estuve tentado de hacerlo, pero al fin de cuentas creo que el marco bidimensional es interesante independientemente.

23. Por supuesto pueden existir casos limítrofes en los cuales es indeterminado si un concepto se referiría a un cierto objeto si un mundo dado resultase ser real. Esto no es ningún problema: podemos permitir indeterminaciones en una intensión primaria, así como a veces permitimos indeterminaciones en la referencia en nuestro propio mundo. Siempre puede haber casos en los que haya más de un candidato igualmente bueno para el referente de un concepto en un mundo (como tal vez sucede con la "masa" en

un mundo relativista, que podría referirse a la masa en reposo o a la masa relativista); así como toleramos la referencia dividida en esos casos reales, ocasionalmente deberíamos esperar referencias divididas en el valor de una intensión primaria.

Con algunos casos limítrofes puede ocurrir incluso que la consideración de que un objeto pertenezca o no a la extensión de un concepto dependa de diversos factores históricos accidentales. Un artículo estimulante de Wilson (1982) analiza casos como estos incluyendo, por ejemplo, un caso hipotético en el cual los druidas podrían llegar a clasificar a los aviones como "pájaros" si los viesen por primera vez volando encima de sus cabezas, pero no si la primera vez hubiesen encontrado uno que se estrelló en la jungla. Podríamos tratar de clasificar estos dos escenarios diferentes como modos distintos en los que podría resultar el mundo real y, por lo tanto, conservar una intensión primaria fija y detallada; o podríamos considerar casos de este tipo como indeterminados respecto de su intensión primaria nuclear. De cualquier forma, un poco de tolerancia en torno de los límites de una intensión primaria es totalmente compatible con mis aplicaciones del marco conceptual.

24. Véase Field, 1973. Creo que un análisis en términos de intensiones primarias puede proporcionar una forma de considerar que el "cambio de significado" es mucho menos frecuente de lo que por lo general se supone. El caso de la relatividad no nos da ninguna razón para creer que la intensión primaria de "masa" haya cambiado desde el siglo pasado a este, por ejemplo, aunque nuestras creencias acerca del mundo real ciertamente se modificaron. (Por razones vinculadas, podríamos intentar utilizar un análisis en términos de intensiones primarias para resistir al "holismo de significado" acerca del pensamiento.) Es probable que cualquier "desarrollo" en las intensiones primarias sea, en el mejor de los casos, del tipo más sutil sugerido por los ejemplos en Wilson (1982), en los que el núcleo permanece fijo, pero accidentes históricos pueden producir diferencias en las prácticas de clasificación en torno de los límites. Pero todo esto merece un desarrollo mucho más extenso.

25. Algunas diferencias: 1) El contenido de Kaplan corresponde muy estrechamente a una intensión secundaria, pero él presenta el carácter como una función del contexto en el contenido, mientras que una intensión primaria es una función del contexto en la extensión. Dada la rigidificación, sin embargo, una intensión primaria es directamente derivable de un carácter y viceversa. Utilizo la primera por razones de simetría y simpleza. 2) Kaplan utiliza su concepción para tratar con los términos indicativos y demostrativos como "yo" y "ese", pero no la extiende para tratar con términos de tipos naturales como "agua", ya que considera que "agua" selecciona al $H_2O$ en todos los contextos (la palabra con sonido similar en Tierra Gemela es simplemente una palabra diferente), y considera que el proceso de determinación de la referencia es aquí una parte de la "metasemántica", en lugar de la semántica. Como antes, el hecho de que sea parte de la metasemántica o de la semántica tiene poca importancia para nuestros propósitos; todo lo que importa es que la determinación de la referencia depende de cómo resulta el mundo real.

26. Podría parecer que la intensión primaria sólo está bien definida sobre mundos posibles centrados en individuos que piensan un pensamiento apropiado o realizan una emisión apropiada. Creo que la intensión primaria es naturalmente extensible a una clase más amplia de mundos: podemos mantener el concepto de nuestro propio mundo y determinar cómo se aplica a otros mundos considerados reales (véase Chalmers 1994c), aunque en algunos su referencia podría estar indeterminada. Pero esto no hará mucha diferencia en lo que sigue.

27. Nótese que estrictamente hablando la intensión primaria selecciona el líquido en nuestro ambiente *histórico*: si viajo a Tierra Gemela y digo "agua", todavía hago referencia al $H_2O$.

28. La relación entre la segunda y la tercera consideración en este apartado —esto es, entre la revisabilidad empírica de Quine y la necesidad *a posteriori* de Kripke— es compleja e interesante. Como observa Kripke, el marco que él desarrolla da cuenta de algunos pero no de todos los problemas planteados por Quine. El análisis de Kripke da cuenta de las revisiones *a posteriori* de las intensiones y, por lo tanto, de cambios en un cierto sentido del "significado". Sin embargo, el análisis bidimensional concuerda con la concepción de intensión única acerca de los valores de verdad que asigna al mundo real, de modo que no da cuenta de la posibilidad quineana de ciertas supuestas verdades conceptuales *a priori* que resultan ser falsas en el mundo real ante suficiente evidencia empírica. Me parece que estas supuestas verdades conceptuales simplemente no son en absoluto verdades conceptuales, aunque podrían ser aproximaciones cercanas.

29. Un tema sutil que surge al utilizar el marco bidimensional para capturar el contenido del pensamiento es que, a veces, un pensamiento puede avalar un mundo centrado como un ambiente potencial aunque no contenga una copia del propio pensamiento. Por ejemplo, si pienso "Estoy en coma", avalo esos mundos centrados en los cuales el individuo en el centro está en coma, tenga o no pensamientos. De modo que debemos tener cuidado cuando definimos las intensiones primarias y las proposiciones primarias para los pensamientos; más cuidado del que tuvimos en el caso del lenguaje.

30. Los mundos deberían considerarse prelingüísticamente, tal vez como distribuciones de cualidades básicas. Es probable que sea mejor no considerarlos colecciones de enunciados, ya que estos *describen* un mundo, y hemos visto que lo pueden hacer de más de un modo. Considerar un mundo como una colección de enunciados sería perder esta distinción. Tal vez los mundos puedan considerarse colecciones de proposiciones (Adams, 1974), si se entiende a las mismas de un modo apropiado, propiedades maximales (Stalnaker, 1976), estados de cosas (Plantinga, 1976), universales estructurales (Forrest, 1986) u objetos análogos a nuestro propio mundo (Lewis, 1986a). Quizá la noción pueda simplemente considerarse primitiva. En cualquier caso, el discurso de mundos posibles estará tan bien o tan pobremente fundado como el discurso sobre la posibilidad y la necesidad en general. Como sucede con las nociones matemáticas, es posible desplegar estas nociones modales de un modo útil aun antes de que tengamos un análisis ontológico satisfactorio.

Siempre consideraré los mundos de una manera "cualitativa", y haré

abstracción de las cuestiones de "individualidad". Es decir, consideraré que dos mundos cualitativamente idénticos son idénticos y no me ocuparé de cuestiones acerca de si los individuos en esos mundos podrían tener "identidades" diferentes (algunos argumentaron que podrían). Estas cuestiones de la identidad a través de mundos plantean muchos problemas interesantes, pero son básicamente irrelevantes para mi utilización del marco conceptual de los mundos posibles.

31. En particular, algunos podrían negar la equiparación del significado y la intensión en el caso de los términos matemáticos. Suele sostenerse que enunciados como "Existe un número infinito de números primos" no son verdaderos en virtud de su significado, a pesar de que son verdaderos en todos los mundos posibles; puede suponerse que los que hacen esta afirmación se resisten a una equiparación del significado con la intensión.

Otros podrían oponerse a la tesis de que la intensión primaria de un término como "agua" es parte de su significado; quizá piensen que el significado del término se agota en su referencia, y que la intensión primaria es parte de la pragmática más que de la semántica. Aun otros podrían resistirse a la afirmación de que la intensión *secundaria* es parte de su significado. En cualquier caso, nada depende del uso de la palabra "significado". Aquí estamos interesados en la verdad en virtud de la intensión, sean o no significados las intensiones.

32. Esta definición de la conceptibilidad se relaciona con la enunciada por Yablo (1993), según la cual $P$ es concebible si podemos imaginar un mundo que permita verificar $P$. La diferencia es que la cláusula "que permita verificar" de Yablo deja espacio para una descripción errónea de situaciones concebidas, de modo que esta variedad de conceptibilidad es, cuanto más, una guía revocable de la posibilidad. En mi definición se elimina esta fuente de revocabilidad. Por supuesto, reaparece en la forma de una brecha mayor entre lo que encontramos concebible a primera vista y lo que es realmente concebible; de modo que debemos ser más reflexivos al hacer juicios de conceptibilidad.

33. Podemos posiblemente aplicar esta crítica al argumento de Descartes de que como él puede concebirse en estado incorpóreo, entonces es posible que sea incorpóreo, de modo que entonces es una entidad no física (ya que cualquier entidad física es necesariamente corpórea). "Yo soy incorpóreo" puede ser concebible-1 y por lo tanto posible-1, pero no se deduce que sea concebible-2 o posible-2. En contraste, el sentido en el cual "Yo soy corpóreo" sería necesario si él fuese un objeto físico es necesidad -2, no necesidad-1. (La intensión primaria de su concepto "yo" selecciona el individuo en el centro de cualquier mundo; la intensión secundaria selecciona a Descartes en todo mundo.)

34. Podríamos decir que no hay nada especialmente "metafísico" acerca de la necesidad metafísica. Visto de este modo, es meramente una marca de necesidad conceptual con un giro semántico *a posteriori* que surge de la naturaleza bidimensional de nuestros conceptos. Para más detalles sobre el tema de que la necesidad *a posteriori* refleja tanto convención como metafísica, véase Putnam (1983) y Sidelle (1989, 1992).

35. Horgan (1984c) expone y argumenta en favor de la posición de que

todos los hechos de alto nivel supervienen lógicamente a los hechos microfísicos. Tal como él lo formula, esos hechos están ligados a lo microfísico por "restricciones semánticas", de modo que todo lo que hay en el mundo es microfísica y "hermenéutica cósmica". Sin embargo, él evita visiblemente el problema de la experiencia consciente. Otros que proponen versiones de la tesis de superveniencia lógica incluyen a Jackson (1993), Kirk (1974) y Lewis (1994).

36. Para argumentos de que los hechos acerca de entidades abstractas son lógicamente supervenientes a lo físico, véase Armstrong (1982).

37. Posiblemente, la experiencia consciente contribuye a la intensión primaria, si se individualiza el agua en parte según el tipo de experiencia a la que da origen. La indicatividad ciertamente contribuye, como se ve en el "nuestro" de "el liquido bebible y claro en nuestro ambiente". Estos hechos no debilitan la superveniencia lógica módulo la experiencia consciente y la indicatividad.

38. Como es familiar a partir del método del enunciado de Ramsey para la aplicación de términos teóricos; véase Lewis (1972).

39. Un ejemplo propuesto como caso problemático por Ned Block.

40. Jackson (1993) y Lewis (1994) formulan un punto similar acerca del requerimiento de analizabilidad para la superveniencia.

41. Una concepción alternativa sostiene que para que algo sea rojo, debe ser el tipo de cosa que tiende a causar *juicios* de rojo. Esto eliminaría los problemas que aquí analizamos, ya que es plausible que los juicios sean lógicamente supervenientes a lo físico.

42. Excepto, posiblemente, en el hecho de que cuando yo tengo una creencia acerca de Bill Clinton, mi duplicado tiene una creencia acerca del duplicado de Clinton. Como es usual, estas cuestiones acerca de la identidad a través de los mundos pueden dejarse de lado.

43. Lo más cercano a un argumento de este tipo es el formulado por la versión de Wittgenstein enunciada por Kripke (1982), quien argumenta que no puede haber ninguna implicación de los hechos físicos y fenoménicos a los hechos intencionales, ya que la implicación no puede estar mediada por un análisis físico, funcional o fenoménico de los conceptos intencionales. Los argumentos (en particular aquellos en contra de un análisis funcional) son discutibles pero, de cualquier forma, como se hizo notar antes, la conclusión del argumento no es que los hechos intencionales sean hechos ulteriores, sino que estrictamente no son hechos en absoluto.

Si se acepta el argumento de Kripke-Wittgenstein en contra de la implicación, la intencionalidad se encuentra en una posición similar a la de la moralidad, que mencionaremos más abajo. En los dos casos, 1) posiblemente no haya ninguna implicación conceptual de los hechos $A$ a los hechos $B$, pero 2) si hay hechos $B$ en nuestro mundo, entonces son válidos en todo mundo concebible $A$-indiscernible. La única conclusión razonable es que estrictamente hablando no haya ningún hecho $B$, y las atribuciones $B$ deben ser tratadas de algún modo deflacionario. La posibilidad de que los hechos $B$ sean hechos ulteriores fundamentales queda descartada por consideracio-

nes de conceptibilidad que muestran que debe haber un vínculo *a priori* de los hechos *A* a los hechos *B* si los hechos *B* están instanciados.

44. Si existen hechos morales subjetivos, entonces las atribuciones morales tienen condiciones de verdad determinadas, pero estas son dependientes de quien hace la atribución. Si esto es así, entonces los conceptos morales tienen una intensión primaria indicativa, y existe indicatividad módulo superveniencia lógica. Este análisis es avalado por los defensores del "realismo moral subjetivista" (Sayre-McCord, 1989), que interpretan el "bien" como "bien para mí" o "bien según mi comunidad". La subjetividad involucrada hace que esto sea un tipo de realismo muy débil, sin embargo. Por ejemplo, según este enfoque resulta que dos personas que discuten sobre lo que es "bueno" podrían no discrepar en absoluto.

45. Los argumentos de Kripke (1972), como aquellos que conciernen a la referencia de "Gödel" en diversas situaciones, sugieren que la intensión primaria asociada con el uso de un nombre no puede en general resumirse mediante una descripción breve. También pueden sugerir que la intensión primaria no puede resumirse mediante alguna descripción finita, aunque estoy menos seguro de esto (ciertamente, establecen que cualquier descripción de esta clase debe incluir un elemento metalingüístico y una condición que requiere una conexión causal apropiada con el agente). Pero nada en estos argumentos sugiere que un nombre (tal como se lo utiliza en cualquier ocasión particular) carece totalmente de una intensión primaria. Los propios argumentos de Kripke incluyen considerar cómo la referencia de un nombre depende del modo como el mundo real resulta; es decir, evalúa la intensión primaria del nombre en diversos mundos centrados.

46. Para un lúcido análisis sobre esta cuestión, véase Nagel, 1986.

47. ¿Cómo eluden los hechos negativos los argumentos en favor de la superveniencia lógica de más arriba? El argumento a partir de la conceptibilidad falla, como lo muestra el ejemplo de los ángeles. El argumento a partir de la epistemología falla ya que claramente *existe* un problema epistemológico acerca de cómo podemos conocer aseveraciones universales de alcance irrestricto (no podemos estar *seguros* de que no haya ángeles). El argumento a partir de la analizabilidad falla, ya que no existe ningún análisis de estos hechos negativos totalmente en términos de hechos positivos (a menos que incorporemos el hecho de segundo orden "eso es todo").

48. ¿Cómo evitan las leyes los argumentos en favor de la superveniencia lógica de más arriba? El argumento a partir de la conceptibilidad falla, como lo muestra el ejemplo de más arriba. El argumento a partir de la epistemología falla, ya que claramente hay problemas con la epistemología de las leyes y la causalidad, como lo atestigua el desafío escéptico de Hume. El argumento a partir del análisis falla, ya que la legalidad requiere una regularidad universal que sustente contrafácticos, y los contrafácticos pertinentes no pueden analizarse en términos de hechos particulares acerca de una historia de mundo (*pace* Lewis, 1973). Los hechos particulares acerca de la historia espaciotemporal del mundo son compatibles con la verdad de todo tipo de contrafácticos diferentes.

49. Enfoques humeanos de las leyes y la causalidad pueden hallarse en

Lewis (1986b), Mackie (1974) y Skyrms (1980). Para argumentos en contra de estos enfoques, véanse Armstrong (1982), Carroll (1994), Dretske (1977), Molnar (1969) y Tooley (1977).

50. En cambio, entre los que parecen sostener que la superveniencia lógica es la regla y no la excepción se encuentran Armstrong (1982), Horgan (1984c), Jackson (1993), Lewis (1994) y Nagel (1974).

51. Para más sobre esto, véase Horgan y Timmons (1992b).

## Capítulo 3

1. Kirk (1974) proporciona una descripción vívida de un zombi, e incluso formula una situación que podría llevarnos a creer que alguien en el mundo real se volvió un zombi, especificando casos intermedios apropiados. De modo similar, Campbell (1970) analiza un "hombre de imitación" que es físicamente idéntico a una persona normal, pero que carece totalmente de experiencia.

2. Kirk (1974) argumenta en favor de la posibilidad lógica de los zombis de este modo indirecto.

3. Jacoby (1990) sostiene la excelente tesis de que los argumentos de conceptibilidad no plantean más problemas para las concepciones funcionalistas de la conciencia que para las concepciones materialistas en general. Lo Interpreta como un argumento en favor de las concepciones funcionalistas, mientras que yo lo interpreto como un argumento en contra de las concepciones materialistas.

4. En realidad, esto terminará intercambiando el rojo con el amarillo en lugar de con el azul, ya que ambos están en los extremos positivos de sus ejes. Sin embargo, los detalles son inesenciales. Para un lúcido análisis de las complejidades del espacio humano del color, véase Hardin (1988).

5. Hardin (1987, p. 138) acepta este punto. Dice que este tipo de inversión es sólo "estrafalario" y no "conceptualmente incoherente".

6. De un modo similar, Gunderson (1970) habla de una "asimetría investigativa" entre las afirmaciones de primera y tercera persona.

7. Thompson (1992) señala que en una habitación en blanco y negro María puede todavía tener experiencias de color, por ejemplo, cuando se frota los ojos. Para evitar esto, tal vez deberíamos estipular que María padece de ceguera del color desde el nacimiento.

8. Churchland (1995, p. 193) y Dennett (1991, p. 281) invocan el vitalismo en este contexto.

9. Al final de la segunda parte de su libro, Dennett promete que en la tercera parte mostrará por qué su concepción funcional puede explicar todo lo que en la conciencia necesita explicación, pero los argumentos son difíciles de localizar. Gran parte de la exposición consiste en observaciones acerca del procesamiento cognitivo con las que alguien como yo podría estar de acuerdo. La cuestión no es si su concepción del procesamiento es correcta, sino si explica la experiencia. El argumento crucial parece encontrarse en el diálogo en las pp. 362-68, donde afirma (en efecto) que lo que debe explicarse es cómo las cosas *parecen*, y que su teoría explica justamente eso. No obstante,

como argumento en el capítulo 5, Dennett confunde un sentido psicológico y un sentido fenoménico de "parecer". Lo que la teoría podría explicar es nuestra disposición a hacer ciertos *juicios* acerca de los estímulos, pero esos juicios nunca fueron explananda problemáticos.

Hay también algunos argumentos en el capítulo 12: 1) un argumento en contra de la posibilidad empírica de los qualia invertidos (que dejan la conducta constante); pero la imposibilidad empírica aquí es compatible con la posición no reductiva; 2) un argumento en contra del argumento a partir del conocimiento de Jackson; lo analizo en el capítulo 4, 3) una afirmación de que el epifenomenalismo acerca de los qualia es ridículo; analizo esta cuestión en los capítulos 4 y 5. Las cuestiones de posibilidad natural y posibilidad lógica suelen ser utilizados en forma conjunta en el análisis de Dennett. Por ejemplo, el filósofo supone que los "qualófilos" sostendrán que una máquina computacional no tendrá experiencias y, por lo tanto, dedica mucho espacio a argumentar que esas máquinas podrían ser conscientes del modo como nosotros lo somos. Pero esto es totalmente compatible con la posición no reductiva; argumento en favor de la misma tesis en capítulos posteriores.

10. Más recientemente, Crick y Koch comenzaron a indagar más allá de las oscilaciones de 40 hertz en su búsqueda de una base neuronal de la conciencia, pero puede aplicarse consideraciones similares. Las oscilaciones tienen la virtud de proporcionar un ejemplo directo.

11. Citado en *Discover*, Noviembre de 1992, p. 96. Crick (1994, p. 258) también acepta la posibilidad de que la ciencia no pueda explicar los qualia, aunque es más circunspecto.

12. Edelman (1989, p. 168) es claro acerca de esto. "Es suficiente proporcionar un modelo que explique su discriminación, variación y consecuencias. Como científicos, no podemos preocuparnos por los misterios ontológicos que conciernen al *por qué* hay algo y no nada, o *por qué* la tibieza se siente como tibieza". Hace una analogía con la teoría cuántica del campo, que nos da una base para discriminar energías y estados materiales, pero que no nos dice por qué existe la materia en primer lugar. Esta analogía es muy compatible con el enfoque no reductivo que desarrollo en capítulos posteriores.

13. Por cierto, si la conciencia fuese lógicamente superveniente a lo físico, entonces estas interpretaciones de "colapso" no podrán despegar, ya que cualquier justificación en favor del tratamiento especial de la conciencia en las leyes desaparecería.

## Capítulo 4

1. De modo similar, Edelman (1992) subtitula su libro (supuestamente materialista) *How the Mind Originates in the Brain* [Cómo la mente se origina en el cerebro].

2. Según mi interpretación, el enfoque de Searle puede entenderse mucho más naturalmente como dualismo de propiedades que como materialismo, a pesar del propio punto de vista de Searle acerca de la cuestión. La aseveración de que los estados cerebrales causan estados fenoménicos y la utilización de argumentos sobre zombis apoyan esta interpretación, como lo

hace también la afirmación de que "lo que ocurre en el cerebro son procesos neurofisiológicos y conciencia, y nada más". El argumento de Searle acerca de la intencionalidad, en su capítulo 8, también apoya esta interpretación. Searle argumenta que la intencionalidad es real (p. 156), pero que los hechos intencionales no pueden estar constituidos por hechos neurofisiológicos (pp. 157-58). La única solución del problema, argumenta, es que la conciencia debe ser parcialmente constitutiva de la intencionalidad, ya que esta es la única otra cosa en la ontología cerebral. Este argumento parece *presuponer* el dualismo de propiedades acerca de la conciencia.

Al explicar su ontología en el capítulo 5, Searle sostiene que la conciencia es irreducible, pero que esto no tiene consecuencias profundas. Dice que los fenómenos como el calor son reducibles sólo porque los redefinimos para eliminar el aspecto fenoménico (al modo expuesto en el capítulo 2), pero que este tipo de redefinición es trivialmente inaplicable a la conciencia, ya que consiste enteramente en su aspecto subjetivo. Esto parece correcto. Tal como lo formulo en el capítulo 2, fenómenos como el calor son reductivamente explicables sólo módulo la experiencia consciente. Pero, continúa diciendo que "esto muestra que la irreductibilidad de la conciencia es una consecuencia trivial de la pragmática de nuestras prácticas de definición" (p. 122). Esto parece interpretar las cosas al revés. Más bien, las prácticas son consecuencia de la irreductibilidad de la conciencia. ¡Si no hacemos abstracción de la experiencia del calor, no podríamos reducir el calor en absoluto! De esta forma la irreductibilidad es una *fuente*, no una consecuencia, de nuestras prácticas. Es difícil ver cómo algo de esto podría trivializar la irreductibilidad de la conciencia.

3. Jackson (1980, 1994), Lewis (1994) y White (1986) formularon argumentos estrechamente relacionados de por qué un materialista no puede apelar a la necesidad *a posteriori*.

4. Jackson (1980) sostiene un punto similar, argumentando que aunque las consideraciones *a posteriori* pueden establecer la fisicalidad de la propiedad *dolor*, todavía surge un problema para el materialismo debido a la propiedad *presentes dolorosos.*

5. Bealer (1994) también sugiere aquí depender del término físico como estrategia, aunque él no sigue el razonamiento hasta sus conclusiones naturales.

6. Algunos pocos adoptaron explícitamente esta posición en trabajos publicados. La mayoría de los que apelan a la necesidad *a posteriori* en defensa del materialismo recurren a las consideraciones kripkeanas (por ejemplo, Hill, 1991; Lycan, 1995; Tye, 1995), y casi nadie defendió explícitamente el tipo más fuerte de necesidad metafísica con este fin. Sin embargo, en una interpretación natural, Bigelow y Pargetter (1990), Byrne (1993), Levine (1993) y Loar (1990) están implícitamente comprometidos con una posición como esta. Byrne, Levine y Terry Horgan defendieron la posición en comunicaciones personales.

7. A veces escuchamos que las verdades matemáticas son metafísicamente necesarias pero no conceptualmente necesarias. Esto depende de cuestiones sutiles acerca del análisis de los conceptos matemáticos y la

necesidad conceptual pero, no obstante, se acepta en general que las verdades matemáticas son *a priori* (con la leve salvedad que se menciona en el apartado siguiente del texto). Más crucialmente, ni siquiera existe un mundo concebible en el cual las verdades matemáticas sean falsas. De modo que estas verdades no hacen que el espacio de mundos posibles sea más pequeño que el conjunto de mundos concebibles.

Podría sugerirse que la superveniencia moral es un ejemplo de superveniencia metafísica sin una conexión *a priori*, pero el caso en favor de la necesidad metafísica fuerte parece aun más débil aquí que en el caso de la experiencia. Existen opciones disponibles (antirrealismo, conexión *a priori*) que son mucho más agradables para la experiencia consciente que las alternativas correspondientes. Más aún, ni siquiera parece haber un mundo concebible que sea física y mentalmente idéntico al nuestro, pero moralmente distinto. De modo que, una vez más, la superveniencia moral no plantea ninguna nueva restricción sobre el espacio de mundos posibles.

8. Jackson (1995) ofrece un argumento simple que pone de relieve lo extraño que resultan las posiciones fisicalistas que no involucran un vínculo *a priori* de lo físico con la psicológico (esto podría aplicarse igualmente a la posición de "necesidad metafísica fuerte" y a la posición de "limitación cognitiva"):

> Es inverosímil que haya hechos acerca de organismos muy simples que no puedan ser deducidos *a priori* a partir de información suficiente sobre su naturaleza física y cómo interactúan con sus ambientes, descriptos físicamente. La historia física acerca de las amebas y sus interacciones con el ambiente es toda la historia acerca de las amebas... Pero según el materialismo sólo diferimos de las amebas esencialmente en la complejidad de los ingredientes y su organización. Es difícil ver cómo ese tipo de diferencia podría generar hechos importantes acerca de nosotros que en principio desafíen nuestros poderes de deducción ... Piénsese en las gráficas en las clases de biología que muestran la progresión evolutiva desde las criaturas unicelulares en el extremo izquierdo hasta los monos superiores y los seres humanos en el extremo derecho: ¿dónde en esa progresión puede el fisicalista afirmar que es atendible que surja el fracaso de la deducibilidad *a priori* de hechos importantes acerca de nosotros mismos? O, si se reduce a eso, ¿dónde en el desarrollo de cada uno de nosotros, a partir de una cigota, podría el materialista localizar plausiblemente el lugar en el cual surgen hechos importantes acerca de nosotros mismos que no pueden deducirse de nuestra historia física?

9. John O'Leary-Hawthorne y Barry Loewer sugirieron, en forma independiente y en conversaciones personales, la analogía entre la superveniencia psicofísica y las verdades matemáticas complejas.

10. Nótese que este razonamiento proporciona una diferencia con las verdades matemáticas aun para alguien que adopte la posición fuerte de que hay ciertas verdades matemáticas tan profundas que no son cognoscibles *a priori* por *ninguna* clase de seres.

11. De un modo extraño, esta posición es bastante cercana a la de los

reduccionistas como Dennett. Después de todo, ambos sostienen que las intuiciones relevantes surgen del déficit cognitivo. La diferencia principal es que el reduccionista piensa que algunos de nosotros podemos superar este déficit, mientras que el objetante actual sostiene que ninguno de nosotros puede hacerlo. Pero, por todo lo que sabe este objetante, otros podrían ya haber logrado el esclarecimiento (tal vez incluso el propio Dennett). Después de todo, ¡los deficitarios no apreciarían una solución de los iluminados!

12. La versión más explícita del argumento a partir de la posibilidad lógica lo enuncia Kirk (1974). También está presente en Campbell (1970), Nagel (1974), Robinson (1976) y otros. Mi presentación del argumento difiere principalmente en la utilización de la noción de superveniencia para proporcionar un marco unificador y en la consideración del papel de la necesidad *a posteriori*. Para un argumento relacionado a partir de la posibilidad del espectro invertido, véase también Seager (1991).

13. Parece haber un sentido razonable en el cual "El agua es acuosa" y "El $H_2O$ es acuoso" expresan diferentes hechos, al igual que "El agua es $H_2O$" y "El $H_2O$ es $H_2O$". En este sentido, individualizamos los hechos por las intensiones primarias de los términos utilizados para expresarlos, en lugar de por las intensiones secundarias.

14. Lockwood (1989, pp. 136-37) sostiene esencialmente este punto. Como lo señala, el que no sepamos que es el mismo hecho que corresponde a cada modo de presentación debe atribuirse a nuestra ignorancia de algún hecho o hechos sustantivos ulteriores, bajo cualquier modo de presentación. Conee (1985a) sostiene una tesis relacionada.

15. Loar sugiere que los conceptos fenoménicos son conceptos *de reconocimiento*, y argumenta que es razonable esperar que un concepto de reconocimiento *R* "introduzca" la misma propiedad que una propiedad teóricamente especificada *P*. Da el ejemplo de alguien que es capaz de reconocer ciertos cactus en el desierto de California sin tener ningún conocimiento teórico acerca de ellos. Pero esto parece erróneo: si el sujeto no puede saber que *R* es *P a priori*, entonces la referencia a R y P se determina de modos diferentes y las intenciones que permiten determinar la referencia pueden disociarse en ciertas situaciones concebibles. A menos que invoquemos la maquinaria adicional de la necesidad metafísica fuerte, la diferencia en las intensiones primarias corresponderá a una diferencia en las propiedades de determinación de la referencia.

En un punto Loar sugiere que los conceptos de reconocimiento refieren "directamente" sin la ayuda de propiedades de determinación de la referencia (intensiones primarias), pero esto también parece erróneo. El propio hecho de que un concepto *podría* referir a alguna otra cosa (un conjunto diferente de cactus, digamos) en una situación concebible diferente nos dice que se encuentra involucrada una intensión primaria sustancial. De esta manera, la referencia no puede ser verdaderamente "directa" en el sentido pertinente.

16. Alguien muy interesado por el problema de la indicatividad podría sostener que la localización del centro de un mundo centrado puede tener significación ontológica o, tal vez, que podría haber una diferencia en los hechos indicativos entre los mundos posibles ordinarios (¿si existiese, quizás,

algo como el "sí mismo objetivo" de Nagel?). Estas cuestiones, al igual que la condición ontológica de la indicatividad en general, me resultan bastante oscuras. La exposición en el texto se basa en el supuesto del oponente de que la indicatividad no lleva a una brecha ontológica, para notar que aun así no puede sostenerse la analogía con la brecha en el caso fenoménico.

17. Existe otro modo como los pensamientos acerca de la experiencia pueden verdaderamente ser como las expresiones indicadoras, a saber, cuando seleccionamos una experiencia como "*esta* experiencia". Cuando hacemos referencia a una de dos experiencias cualitativamente idénticas (como en el escenario de "Dos Tubos" de Austin, 1990), es concebible que el conocimiento de todos los hechos "objetivos" pueda dejar indeterminada la cuestión de a qué experiencia se hace referencia. (Nótese que la cuestión aquí concierne a la referencia de ejemplares, no de tipos.) En caso de que un materialista pudiese querer utilizar este problema para obtener apoyo en contra del argumento a partir del conocimiento, hago notar que 1) en este caso el hecho epistémico ulterior es independiente de los hechos fenoménicos (aun conocer todos los hechos fenoménicos no nos dice cuál es *esta* experiencia); 2) no proporciona una situación en la cual existe un mundo concebible *no centrado* que difiere de este mundo, de modo que no se lo puede utilizar para construir un argumento ontológico como el del texto, 3) cuanto más, el hecho ulterior sirve para *localizar* qué ocurre en el centro de un mundo, al decirnos qué entidad es *esta* al modo como los hechos indicativos nos dicen qué entidad soy *yo*.

La verdadera moraleja aquí es que en ciertas circunstancias, necesitamos colocar más información en el centro de un mundo: no sólo señalar un individuo y un tiempo (como "yo" y "aquí"), sino también señalar una experiencia (como "esto"). Algo similar se aplica posiblemente a los demostrativos de orientación, tales como "izquierda" y "derecha" (conocer los hechos objetivos acerca de un mundo podría no decirnos cuál dirección es izquierda y cuál es derecha). Todo esto está asociado con brechas epistémicas de la variedad indicativa relativamente no amenazadora: en ninguno de estos casos existen mundos *no centrados* en los cuales los hechos básicos son válidos pero los hechos ulteriores no.

18. En una objeción relacionada, Churchland (1985) sugiere que el argumento de Jackson utiliza ambiguamente el término "conocer": María tiene un conocimiento *proposicional* o *sentencial* completo de los hechos físicos, pero carece de conocimiento *por familiaridad* con las experiencias de rojo. La respuesta es similar. En tanto el conocimiento de la experiencia de rojo de María *constriña el modo como el mundo es* será conocimiento fáctico y el argumento es exitoso (no se requiere ninguna aseveración de que el conocimiento fáctico deba ser "sentencial"). De manera que, al igual que Lewis y Nemirow, Churchland está comprometido con la aseveración poco razonable de que el conocimiento de la experiencia de rojo de María no le dice nada acerca del modo como el mundo es.

19. Lycan (1995) ofrece nueve (!) argumentos al efecto de que el conocimiento de María involucra nueva información.

20. Existen algunas otras respuestas al argumento a partir del cono-

cimiento que no analicé, pero mi respuesta a ellas debería ser predecible. Para mencionar sólo una: Dretske (1995) sostiene que el conocimiento del cual carece María es conocimiento acerca de su ambiente. Si ella supiese más sobre la composición de las cosas rojas, sabría qué representan las experiencias de rojo y así (según la teoría de Dretske) sabría cómo son las experiencias de rojo. Extrañamente, Dretske no considera la objeción obvia: aunque María supiese todo acerca de la composición de los objetos rojos, ¡*todavía* no sabría cómo es ver rojo!

21. Agradezco a Frank Jackson la discusión sobre este punto.

22. El propio Kripke acepta (1972, f.74) que la mera ausencia de identidad puede ser una conclusión débil. Pero hace notar que los argumentos modales podrían también utilizarse en contra de formas más generales de materialismo.

23. De un modo similar, los argumentos a partir de la incorporeidad podrían establecer que las propiedades mentales no son idénticas a las propiedades físicas, en el sentido de que las propiedades físicas sólo pueden ser instanciadas por objetos físicos; Bealer (1994) formula un argumento de esta clase. Pero, nuevamente, este tipo de no identidad es una conclusión débil: todavía es compatible con la superveniencia lógica y, por lo tanto, con el materialismo. Un tipo similar de argumento de no identidad podría utilizarse para casi cualquier propiedad de alto nivel.

24. En su cuidadoso análisis, Boyd (1980, p. 98) hace notar que la posibilidad de los zombis, a diferencia de la posibilidad de la incorporeidad, implica la falsedad del materialismo. En consecuencia proporciona un argumento independiente en contra de esa posibilidad, pero este es esquemático y no resulta convincente. Boyd hace una analogía con un ordenador que computa una función particular, argumentando que 1) puede *parecernos* que podríamos tener todos los circuitos del ordenador tal como son sin que la función sea computada, pero que no obstante esto es imposible, y 2) la aparente posibilidad de los zombis es análoga a esto. Sin embargo, la analogía falla. La situación con el ordenador es análoga a la "posibilidad aparente" (muy tenue) de que pueda haber una réplica física de mí mismo que no aprenda lo que yo aprendo, o que no discrimine lo que yo discrimino. Nada en esta analogía puede dar cuenta de la naturaleza mucho más convincente de la posibilidad aparente de una réplica sin experiencia consciente.

25. La observación de Kripke de que el materialista debe mostrar que "estas cosas que podemos imaginar no son, de hecho, cosas que podamos imaginar" (en el penúltimo párrafo de Kripke, 1971) también sugiere el tratamiento débil. Kripke deja abierta la cuestión de que la aparente posibilidad podría explicarse de algún modo bastante diferente de los casos estándar agua/$H_2O$, pero dice que "debería ser un argumento más profundo y sutil de lo que yo puedo comprender y más sutil de lo que haya aparecido alguna vez en toda la literatura materialista que he leído".

26. Aunque con frecuencia se interpreta que el argumento es una aplicación de la teoría de la designación rígida de Kripke, una versión de ella podría, en principio, haber sido formulada diez años antes de que la teoría fuera desarrollada. Podríamos haber preguntado a los teóricos originales de

la identidad por qué los hechos físicos acerca del $H_2O$ requieren que sea agua (o acuosa), mientras que los hechos físicos acerca de los estados cerebrales no parecen requerir que haya dolor.

27. Horgan (1987) habla de superveniencia "metafísica" en este contexto, como lo hace Byrne (1993). Sin embargo, si tengo razón en que la posibilidad metafísica y la posibilidad lógica (de mundos) coinciden, entonces se deduce la superveniencia lógica.

28. Para otras versiones de este tema, véanse Blackburn (1990), Feigl (1958), Lockwood (1989), Maxwell (1978) y Robinson (1982).

29. Shoemaker (1980) formula una perspectiva del mundo como puro flujo causal, y argumenta que todas las propiedades son "potencias", sin ninguna otra propiedad subyacente a esas potencias. El argumento de Shoemaker en favor de este punto de vista es fundamentalmente verificacionista y no confronta directamente los problemas que el enfoque genera.

Shoemaker sostiene, además, que como las potencias asociadas a una propiedad son esenciales a ella, las leyes de la naturaleza deben ser necesarias *a posteriori* (Swoyer [1982] sostiene algo similar y Kripke [1980] juguetea con la conclusión). El análisis bidimensional de la necesidad *a posteriori* sugiere que debe haber algo malo con esta sugerencia o, al menos, que es más limitado de lo que parece. Cuanto más, podría ocurrir que mundos con leyes diferentes puedan no describirse correctamente como conteniendo (digamos) electrones; estas consideraciones no pueden decretar que esos mundos sean imposibles. Más aún, parece inverosímil sostener que *todas* las potencias asociadas con los electrones sean constitutivas de la "electronicidad". Es más plausible que para que una entidad califique como electrón sólo se requieran algunas de esas potencias, y que mundos moderadamente contranomológicos que contengan electrones sean posibles. Shoemaker sostiene que no hay ninguna manera de distinguir potencias constitutivas de potencias no constitutivas, pero el análisis bidimensional sugiere que esta distinción surge del concepto de electronicidad.

Es necesario distinguir un cierto número de cuestiones. 1) ¿La referencia a las propiedades físicas se fija en forma relacional? (Shoemaker, Chalmers: Sí). 2) ¿Las propiedades físicas son idénticas a las propiedades relacionales (en intensión secundaria)? (S: Sí; C: Probablemente, pero las intuiciones semánticas pueden diferir). 3) ¿Son todas las relaciones nomológicas de una propiedad física esenciales para ella? (S: Sí; C: No). 4) ¿Hay propiedades intrínsecas subyacentes a estas propiedades relacionales? (S: No; C: Sí).

30. Este enfoque ha sido defendido en años recientes por Lockwood (1989) y Maxwell (1978), los cuales presentan este punto de vista como una versión no ortodoxa de la teoría de la identidad. Este enfoque me fue implacablemente impuesto por Gregg Rosenberg.

31. Sin embargo, véase Lahav y Shanks (1992) para un enfoque contrario.

32. Lewis (1990) llega a una conclusión similar de un modo diferente.

33. Esta es la cuestión acerca de la cual de vez en cuando realicé encuestas en ocasión de pronunciar conferencias acerca de la conciencia, y en

otras oportunidades. Los resultados son consistentemente 2:1 o 3:1 en favor de que haya algo ulterior que necesita explicación. Por supuesto, no se trata de que la filosofía se haga mejor mediante la democracia, pero cuando nos ocupamos de una de estas cuestiones que la argumentación no puede resolver, el equilibrio de la intuición previa tiene un cierto peso.

34. *Materialismo biológico.* Un punto de vista común (Hill, 1991; Searle, 1992) es que la conciencia es necesariamente biológica. Según este enfoque, el materialismo es verdadero, pero los sistemas inconscientes con la misma organización funcional que los sistemas conscientes son lógicamente posibles y es probable que también sean empíricamente posibles. Sin embargo, una vez que admitimos la posibilidad lógica de un isomorfo funcional inconsciente de mí mismo, debemos seguramente admitir la posibilidad lógica de un isomorfo *biológico* inconsciente, ya que no existe un vínculo conceptual de la neurofisiología con la experiencia consciente, no más que con el silicio. Por lo tanto, es probable que sea mejor considerar este enfoque como una versión del dualismo de propiedades, en el que la conciencia es un hecho ulterior más allá de los hechos físicos. Si no, entonces, en el mejor de los casos debe combinarse con una apelación a la necesidad metafísica fuerte para sustentar el vínculo entre la bioquímica y la conciencia; esto heredaría todos los problemas de esa perspectiva.

(Searle [1992] admite la posibilidad lógica de los zombis y, de hecho, sostiene que sólo hay una conexión causal entre lo microfísico y la experiencia consciente, de modo que quizá sea mejor considerarlo un dualista de propiedades. Hill [1991] intenta evitar la posibilidad de los zombis apelando a designadores rígidos, pero hemos visto que esta estrategia no es útil.)

35. *Funcionalismo fisicalista.* Según este enfoque popular (por ejemplo, Shoemaker, 1982), la propiedad de tener una experiencia consciente es una propiedad funcional, pero la de tener una experiencia consciente *específica* (una sensación de rojo, digamos) es una propiedad neurofisiológica. Según este punto de vista, el espectro invertido entre isomorfos funcionales es lógica y tal vez empíricamente posible, pero los isomorfos funcionales totalmente inconscientes no lo son. Pero, de nuevo, una vez que aceptamos que un isomorfo funcional invertido es lógicamente posible, debemos también aceptar que un isomorfo *físico* invertido también lo es, ya que la neurofisiología no ofrece una conexión conceptual con una experiencia particular; no más de lo que el silicio lo hace. De manera que, una vez más, parece que los hechos físicos no determinan todos los hechos, y se deduce entonces algún tipo de dualismo de propiedades. Nuevamente, el fisicalismo puede mantenerse sólo si acepta la noción problemática de la necesidad metafísica fuerte.

Esta perspectiva suele formularse como una identificación *a posteriori* de las propiedades fenoménicas con las propiedades neurofisiológicas. Como tal, es vulnerable a los problemas usuales de esta identificación *a posteriori* (¿cuál es la intensión primaria?) además de al argumento de más arriba. Como White (1986) hace notar en una crítica similar, sería mejor que los que defienden este enfoque adhiriesen a un funcionalismo generalizado.

36. *Psicofuncionalismo.* Según este punto de vista, las propiedades mentales se identifican con las propiedades funcionales *a posteriori*, sobre la

base de sus papeles en una psicología empírica madura (véase Block, 1980). Si este enfoque se aplicase a propiedades fenoménicas, las nociones fenoménicas tendrían las mismas intensiones secundarias que las nociones funcionales, a pesar de una diferencia en la intensión primaria. Los problemas con esta posición pueden recibir un mejor análisis según las líneas sugeridas en el apartado 2; es decir, concentrándonos en las intensiones primarias. Si la intensión primaria de las nociones fenoménicas es ella misma funcional, entonces la posición resulta, después de todo, avalada por algún tipo de funcionalismo analítico; pero, si no lo es, entonces el concentrarnos en la propiedad introducida por esa intensión nos llevará invariablemente a una forma de dualismo. De cualquier manera, este enfoque ya no funciona para salvar al materialismo.

Los defensores de esta perspectiva con frecuencia ignoraron el papel de los conceptos en la determinación de la referencia por medio de las intensiones primarias. Aun dada una teoría científica con la "creencia" como término teórico, habrá una historia conceptual que contar acerca de por qué *ese* tipo de estado califica como una creencia, y no como un deseo o cualquier otra cosa. Más probablemente, esta intensión determinante de la referencia será funcional; seleccionará algo así como el estado que desempeñe el papel más creíble de creencia dentro de la teoría, donde "creíble" se traduce según nuestro concepto previo. Cualquiera sea la naturaleza de las intensiones primarias para las propiedades fenoménicas, surgirán problemas allí. Concentrarse en las intensiones secundarias es simplemente barrer los problemas debajo de la alfombra.

Otro problema con el psicofuncionalismo es el siguiente. Implica una especie de chauvinismo, debido a que le asigna un peso extra a la psicología humana cuando decide qué es una creencia, por ejemplo. Véase Shoemaker (1981) para una excelente crítica, pero véase también Clark (1986) para una respuesta. Resulta más plausible que para la mayoría de las nociones mentales, las intensiones primaria y secundaria coincidan. Si no, tendríamos situaciones en las que nosotros y nuestras contrapartes en Tierra Gemela le asignaríamos significados distintos a "creencia", a pesar de que nuestros conceptos previos son idénticos.

37. *Monismo anómalo.* Según este enfoque, cada estado mental es idéntico como ejemplar a un estado físico, pero no hay leyes psicofísicas estrictas. El monismo anómalo fue formulado por Davidson (1970) como una concepción de estados intencionales en lugar de estados fenoménicos, pero se lo puede considerar pertinente por dos razones: primero, ofrece un argumento *a priori* en favor del fisicalismo basándose simplemente en la interacción causal (incluso una interacción en un solo sentido) entre estados físicos y estados mentales y, segundo, niega las leyes psicofísicas que mi enfoque requiere.

Para ver que mi posición no está amenazada por los argumentos de Davidson, nótese que nada en ellos contradice la existencia de leyes *puntuales* de la forma "Si un sistema está en el estado físico máximamente especificado *P*, entonces está en el estado mental (máximamente especificado) *M*". Davidson avala la superveniencia de lo mental a lo físico, lo que, sobre

la base de una interpretación natural (véase Kim [1985] para un análisis), parece tener como consecuencia la existencia de leyes de esa clase. Podría interpretarse que Davidson no niega las leyes puntuales sino las leyes *"de tipo"* más interesantes que conectan estados mentales con estados físicos bajo tipos amplios como los de la psicología popular. Esto es lo más que parece surgir de sus argumentos en favor del holismo de lo mental. De ser así, la superveniencia natural no está amenazada. También se deduce que el argumento en favor de la identidad de ejemplares no funciona. Este se basaba en que *no* existían leyes estrictas para sustentar una conexión causal entre lo físico y lo mental (de modo que era necesario, en cambio, una identidad). Pero aun una ley puntual estricta es suficiente para suscribir el tipo de conexión que yo avalo, de los estados físicos a los estados fenoménicos. De modo que el dualismo tampoco está amenazado.

38. *Representacionalismo.* Una posición recientemente popular (por ejemplo, Dretske, 1995; Harman, 1990; Lycan, 1996; Tye, 1995) es que las propiedades fenoménicas son sólo propiedades *representacionales*, de modo que los qualia de amarillo son meramente estados perceptuales que representan cosas amarillas o algo similar. Por supuesto, la interpretación de esta sugerencia depende de qué concepción se adopte acerca de las propiedades representacionales. Muy frecuentemente, se combina esta sugerencia con una concepción reductiva de la representación (por lo general una concepción funcional o teleofuncional), en cuyo caso se vuelve una variante del funcionalismo reductivo y enfrenta los problemas usuales. Una concepción no reductiva de la representación permite evitar estos problemas (aunque podría tener otros), pero llevaría a una concepción no reductiva de la experiencia.

La plausibilidad superficial de algunas concepciones representacionalistas bien podría surgir de un deslizamiento entre las interpretaciones inflacionaria y deflacionaria de "representación", donde la segunda es una noción puramente funcional (o teleofuncional), pero la primera no lo es. El vínculo entre la fenomenología y la representación se vuelve plausible en la primera interpretación, pero la reducción de la representación se hace plausible en la segunda. Alternativamente, puede invocarse la necesidad metafísica fuerte para establecer la conexión entre estados representacionales y estados fenoménicos, con los problemas asociados. (Entre los representacionalistas contemporáneos, Dretske [1995] y Harman [1990] parecen avalar una posición de tipo *A* fuertemente reductiva, mientras que Lycan [1996] y Tye [1995]) parecen avalar una posición de tipo *B* que se basa en la necesidad *a posteriori.*)

Otro modo de enfocar el representacionalismo es advertir que casi todos sus defensores están de acuerdo en que no *todos* los estados representacionales son estados fenoménicos (aquellos que discrepan casi seguramente son no reductivistas acerca de ambos), de modo que podemos preguntar: ¿Qué es lo que hace que algunos estados representacionales sean estados *fenoménicos*? Es este criterio ulterior el que verdaderamente hace el trabajo en una teoría representacionalista de la conciencia. Con frecuencia, el criterio será algo similar al requerimiento de que el estado representacional esté disponible

para los procesos centrales de un modo apropiado, en cuyo caso es evidente que se trata de una concepción funcionalista reductiva con los problemas usuales (¿por qué *eso* haría que un estado representacional sea fenoménico?). La alternativa es individualizar los estados relevantes como aquellos estados representacionales que son *fenoménicos* pero, entonces, el camino llevaría directamente de vuelta al dualismo de propiedades.

39. *La conciencia como pensamiento de orden superior*. La propuesta de que un estado consciente es un objeto de un pensamiento de orden superior (véase, por ejemplo, Rosenthal [1996], entre otros) puede encararse de un modo similar. Si esto se combina con un enfoque reductivo acerca de qué es tener un pensamiento de orden superior, resultará esencialmente un punto de vista funcionalista reductivo con los problemas usuales. Si no, entonces llevará a un punto de vista no reductivo de la experiencia (tipo *B* o tipo *C*) y, de ese modo, será compatible con el dualismo de propiedades que sugiero, aunque podría tener otros problemas (como analizo en el capítulo 6).

40. *Teleofuncionalismo reductivo.* Vale la pena mencionar el enfoque de Dretske (1995), según el cual un componente teleológico está también incluido en los criterios que definen qué es tener una experiencia. Para tener experiencias, un sistema no sólo debe funcionar de un cierto modo, sino que los procesos relevantes deben haber sido seleccionados de manera apropiada en su historia. Se dice que esta posición es capaz de evitar algunos de los problemas del funcionalismo estándar, en el sentido de que, por ejemplo, toma en consideración (y explica) la posibilidad de zombis funcionalmente idénticos: estos sólo son sistemas con la historia equivocada. Pero padece sus propias versiones de los principales problemas. Por ejemplo, no parece menos lógicamente posible que un sistema funcionalmente idéntico con la historia apropiada pueda carecer de conciencia; del mismo modo, el conocimiento de la organización más la historia no alcanza para darnos conocimiento de la experiencia. Podríamos decir que este enfoque "evita" los problemas del funcionalismo reductivo de un modo equivocado. Finalmente, esta posición está más cerca de un enfoque funcionalista reductivo de tipo *A* que de un enfoque que toma en serio la conciencia.

41. *Causalidad emergente*. Muchos quisieron rechazar una concepción reductiva de la conciencia al mismo tiempo que darle un papel causal central. Un modo popular de hacerlo es argumentar en favor de la causalidad emergente: la existencia de nuevos tipos de causalidad en los sistemas físicos de una cierta complejidad. Por ejemplo, Sperry (1969, 1992) argumentó que la conciencia es una propiedad emergente de sistemas complejos que a su vez desempeña un papel causal; los emergentistas británicos como Alexander (1920) sostuvieron un punto de vista parecido (véase McLaughlin [1992] para un análisis). De manera similar, Sellars (1978; véase también Meehl y Sellars, 1958) sugirió que nuevas leyes de la causalidad física podrían entrar en juego en ciertos sistemas, tales como aquellos hechos de protoplasma o los que sustentan seres sensibles. (El llamó a este enfoque "$\text{fisicalismo}_1$", en oposición al "$\text{fisicalismo}_2$" en el cual los principios físicos básicos que se encuentran en la materia inorgánica se aplican en forma generalizada.) Estos puntos de vista no deben confundirse con el enfoque

"inocente" de la causalidad emergente que puede encontrarse en la teoría de los sistemas complejos y según el cual las leyes de bajo nivel producen conductas cualitativamente nuevas gracias a los efectos de la interacción. Según el enfoque más radical, surgen nuevos principios fundamentales que no son consecuencias de las leyes de bajo nivel.

Existen dos problemas con esta perspectiva. Primero, no hay evidencia de esos principios emergentes de la causalidad. Hasta donde podemos decir, toda causalidad es una consecuencia de la causalidad física de bajo nivel, y la "causalidad descendente" nunca interfiere con los asuntos de bajo nivel. Segundo y tal vez más importante: en un análisis detallado, el enfoque deja a la conciencia tan superflua como antes. Para ver esto, nótese que nada en la historia acerca de la causalidad emergente requiere que invoquemos propiedades *fenoménicas* en algún lugar. La historia causal completa puede contarse en términos de vínculos entre configuraciones de propiedades físicas. Todavía existirá un mundo posible que es físicamente idéntico pero que carece totalmente de conciencia. Se deduce que, en el mejor de los casos, las propiedades fenoménicas se *correlacionan* con configuraciones causalmente eficaces. Si existe algún modo de considerar que las propiedades fenoménicas son causalmente eficaces según este enfoque, la misma maniobra se aplicará a mi perspectiva. De hecho, parece mejor considerar esta postura como una versión de mi propio punto de vista, donde la conciencia superviene a lo físico mediante un vínculo nomológico contingente. Se la puede modificar mediante la adición de nuevas leyes de causalidad física emergente, pero estas simplemente complican las cosas, en lugar de modificar algo fundamental.

42. *Misterianismo.* Los que no simpatizan con las concepciones reductivas de la conciencia suelen sostener que esta última seguirá siendo un eterno misterio. Un punto de vista de esta clase fue examinado por Nagel (1974) y Jackson (1982), y desarrollado por McGinn (1991). Según este enfoque, la conciencia podría estar tanto más allá de nuestra comprensión como el conocimiento de la astronomía lo está de las babosas.

Una perspectiva de esta clase puede ser tentadora, pero es prematura. Decir que no existe ninguna explicación reductiva de la conciencia no significa que no exista ninguna explicación. En particular, una concepción de los principios en virtud de los cuales la conciencia superviene naturalmente a lo físico podría proporcionar una teoría esclarecedora de la conciencia aun en un enfoque no reductivo.

McGinn (1991) argumenta que existe una conexión necesaria entre los estados cerebrales y los estados conscientes (si no el surgimiento de la conciencia sería un milagro), pero que nunca podremos conocer cuál es la conexión. Su análisis sugiere que tiene en mente la necesidad lógica o metafísica; pero el argumento establece cuanto más la necesidad natural. Ciertamente, una conexión nomológica contingente entre la conciencia y lo físico no es más milagrosa que cualquier ley contingente y, seguramente, una conexión de este tipo parece mucho menos misteriosa que una conexión lógica o metafísicamente necesaria que está más allá de nuestra comprensión. Tampoco es obvio por qué no podríamos utilizar nuestro conocimiento de las regularidades que conectan los procesos físicos y la experiencia para inferir

esas leyes. En los próximos capítulos exploraré la caracterización de leyes apropiadas. De esta manera, podemos ver que un enfoque no reductivo de la conciencia no necesariamente debe conducir al pesimismo.

## Capítulo 5

1. Elitzur da crédito de su análisis a Penrose (1987).

2. Dejo de lado aquí las experiencias religiosas. Es posible que lo que verdaderamente debe explicarse aquí sean las experiencias espirituales profundas.

3. No estoy seguro de que esta línea argumental se encuentre explícitamente en la literatura, pero existen argumentos relacionados. Por ejemplo, Foss (1989) responde al argumento a partir del conocimiento de Jackson (1982) haciendo notar que María podría saber todo lo que un sujeto con visión en color *diría* acerca de diversos colores e incluso todo lo que un sujeto *podría* decir. Pero, por supuesto, esto no alcanza en absoluto para saber todo lo que se puede saber.

4. Desde ya, esto no sería realmente útil.

5. Dretske (1995) formula un tipo similar de argumento; sostiene que esta teoría explica el modo como las cosas parecen ser y de esa forma explica lo que debe ser explicado. Una vez más, existe una confusión entre el sentido psicológico y fenoménico de "parecer". Por lo general "parecer" no es un buen término para caracterizar los explananda de una teoría de la conciencia, precisamente debido a esa ambigüedad. Es interesante que usualmente sólo lo utilizan los defensores de concepciones reductivas.

6. Por supuesto, existieron varios ataques a la idea de lo "Dado" y a la idea de "datos sensibles" en la literatura. Pero no creo que ninguno de estos pueda tener éxito en derribar la idea de que tener una experiencia proporciona una fuente de justificación para una creencia acerca de la experiencia. Esos ataques ofrecen buenas razones para rechazar diversas aseveraciones más fuertes, como la afirmación de que todo conocimiento se deriva del conocimiento de la experiencia, que percibimos el mundo percibiendo datos sensoriales, o que tener una experiencia significa automáticamente ponerla bajo un concepto. Sin embargo, yo no sostengo ninguna de esas aseveraciones.

Sellars (1956) critica de un modo plausible la idea de que "experimentar un contenido sensorial *s*" implica conocimiento no inferencial de *s*. Observa que el conocimiento es un estado *conceptual*, de manera que es improbable que ese tipo de conocimiento sea primitivo; pero la experiencia parece ser más primitiva. Tener una experiencia es posiblemente un estado *no conceptual*, y nuestra familiaridad con la experiencia es una relación no conceptual (aunque esta cuestión depende de cómo definimos qué es ser "conceptual"). La cuestión residual es, por lo tanto, la de cómo un estado no conceptual puede proporcionar evidencia de un estado conceptual. Este es un problema difícil, pero no es un problema que sólo se le presente al no reduccionista acerca de la conciencia. Surge, incluso, en el caso del conocimiento perceptual estándar del mundo, en el cual aun un reduccionista debe aceptar que la justificación de una creencia se basa parcialmente en una

fuente no conceptual, a menos que estemos dispuestos a aceptar alternativas que parecen enfrentar dificultades todavía mayores. Pienso que podría formularse una concepción de esa justificación, pero este es un proyecto separado de una extensión considerable. Aquí simplemente hago notar que el no reduccionismo acerca de la conciencia no plantea ninguna preocupación *especial* en esta área.

7. Esta línea argumental es adoptada por Hill (1991) en su detallada respuesta a los argumentos escépticos concernientes a la experiencia basada en la posibilidad del "dolor sustituto", en particular al argumento de Shoemaker (1975a). Hill hace algunas consideraciones que son compatibles con mi tratamiento de estas cuestiones, aunque él defiende un materialismo biológico de tipo B en lugar de un dualismo de propiedades. Estas incluyen una comparación entre argumentos escépticos acerca de la experiencia y argumentos escépticos acerca del mundo externo, y argumentos en contra de una "condición de discernibilidad" que sostiene que no estamos justificados en creer que *P* a menos que en toda situación en la cual carecemos de evidencia de *P* podamos reconocer que carecemos de evidencia de *P*.

8. Esta sugerencia fue hecha por John O'Leary-Hawthorne durante una discusión.

9. Podría pensarse que la línea opuesta - en la cual la creencia de un zombi "Soy consciente" resulta verdadera porque su concepto selecciona una propiedad funcional - podría ser útil para enfrentar los problemas epistemológicos del dualista de propiedades, dado que ya no se deduciría que yo tengo creencias que están justificadas mientras que las de un zombi no lo estarían. Sin embargo, cualquiera sea la línea que adoptemos aquí, el zombi seguirá teniendo *algunas* falsas creencias, como la de que tiene propiedades más allá de sus propiedades físicas y funcionales; el problema de justificar mis creencias correspondientes se reiterará entonces en esta forma.

10. Por ejemplo, por Bill Lycan en una comunicación personal. También podemos retroceder a un concepto como "Experiencia de la clase típicamente causada (en la mayoría de nosotros) por cosas rojas", aunque aquí eludimos el relativismo al costo de la posibilidad de estar sistemáticamente equivocados acerca de la categoría de nuestras propias experiencias.

11. Este tipo de relativismo no ocurre con los conceptos *externos* del color, tal como la rojez como propiedad de *objetos* en lugar de experiencias. En una primera aproximación, la referencia se fija en las cosas rojas como cosas que típicamente dan origen (en la mayoría de nosotros) al mismo tipo de experiencias de color que algunos ejemplos paradigmáticos. Esto es más "público" de dos modos: debido a la referencia a ejemplos paradigmáticos públicos y debido a la referencia a experiencias a través de una comunidad. De esta manera, alguien con un espectro invertido usaría "cosas rojas" para referir a las mismas cosas que yo, aunque su término "experiencias de rojo" seleccione algo diferente.

Podríamos preguntarnos acerca de individuos con fronteras diferentes en su espacio de color, por ejemplo, alguien que piense que las zanahorias tienen el mismo color que las rosas y los tomates. Parece muy natural decir que su emisión de "Las zanahorias son rojas" es falsa, debido al elemento

público en el concepto "rojo", pero tal vez exista un sentido menos público en el que puede considerarse que es verdadera (donde "rojo" significa "rojo para mí"). Aun aquí, sin embargo, no hay *mucho* espacio para el relativismo, porque para cualquier individuo el término todavía estará ligado a paradigmas externos. Aun utilizando este sentido relativista, el término "rojo" de cualquiera debe seleccionar muchas cosas rojas, y lo mismo ocurre con otros conceptos.

Podríamos también eliminar la dependencia de la experiencia de las caracterizaciones de los conceptos de color externos, caracterizándolos en cambio en términos de *juicios*: de esta forma, las cosas rojas son aquellas que típicamente se juzgan como del mismo color que los ejemplos paradigmáticos. Esto tiene la ventaja de permitir que los zombis hablen verazmente de objetos verdes, como podría ser razonable. Después de todo, pareciera que la similitud intersubjetiva en los juicios, más que la similitud en la experiencia, es todo lo que se requiere para hacer que la referencia de los términos de color funcionen.

12. La distinción entre el concepto cualitativo y el concepto relacional de "experiencia de rojo" está estrechamente vinculada a la distinción de Nida-Rümelin (1995) entre la interpretación "fenoménica" y "no fenoménica" de las atribuciones de creencias, tal como "Mariana cree que el cielo le parece azul a Pedro". La lectura "fenoménica" de Nida-Rümelin atribuye una creencia que involucra el concepto cualitativo pertinente al que realiza la atribución; mientras que la interpretación no fenoménica atribuye una creencia que involucra un concepto relacional. (El concepto relacional en los ejemplos de Nida-Rümelin parece ser algo así como el concepto basado en una comunidad mencionado en la nota 10.)

13. Por supuesto, mi utilización del símbolo "*R*" para el concepto es deliberadamente reminiscente de la "*E*" en el argumento del lenguaje privado de Wittgenstein. Ni siquiera intentaré analizar ese argumento aquí; ese proyecto se dificultaría especialmente debido al hecho de que no existe ninguna interpretación ampliamente aceptada acerca de exactamente cuál es el argumento. Basta decir que cada versión del mismo que he visto se apoya en premisas muy dudosas o se aplica tan fuertemente a los conceptos cotidianos como a los conceptos de la experiencia privada, o ambos.

14. Si hacemos algo de introspección, es notable que hay poco que podamos decir, incluso "decirnos a nosotros mismos", que permita distinguir las experiencias de rojo y de verde aparte de señalar las propiedades relacionales, a pesar de nuestra propia percatación de su rica diferencia intrínseca. Esto podría verse como evidencia ulterior de que las cualidades son inherentes al dominio fenoménico y no se reflejan directamente en el psicológico.

15. En cierta medida, esta línea de pensamiento refleja una línea argumental de Shoemaker (1975): si los espectros invertidos son posibles, ni los estados cualitativos ni las creencias cualitativas pueden definirse funcionalmente; Shoemaker lo formula en términos de designación rígida con determinación relacional de la referencia, de modo que se concentra aquí en las intensiones secundarias.

16. Esta observación es paralela a la de Nida-Rümelin (1995) de que la distinción entre creencias fenoménicas y no fenoménicas no es una instancia ordinaria de la distinción *de re / de dicto*.

17. Nótese que no se trata del simple indicativo "este", cuya intensión primaria es la misma sea la experiencia *R* o *S*, y para el cual es una trivialidad no informativa que esta experiencia es *este* tipo de experiencia. Más bien, se trata del "*este*" sustancial con una intensión sustancial que selecciona experiencias *S* en todo mundo centrado.

18. Conee (1985b) se basa en este tipo de relación constitutiva entre los qualia y las creencias cualitativas en su respuesta al argumento epistemológico de Shoemaker (1975a).

## Capítulo 6

1. Nótese que para que estos principios proporcionen leyes psicofísicas, debemos interpretar los juicios de segundo orden como "Estoy teniendo una experiencia de rojo" mediante la interpretación relacional que analizamos en el apartado final del capítulo 5; es decir, de la forma "Estoy teniendo el tipo de experiencia usualmente causada por objetos rojos". Los elementos relacionales del concepto "experiencia de rojo", a diferencia de los elementos cualitativos intrínsecos, se reflejarán en el procesamiento físico: la creencia correspondiente que involucra el concepto cualitativo de "experiencia de rojo" no supervendrá lógicamente a lo físico, de modo que el hecho de que esas creencias sean correctas no proporcionará una ley psicofísica. Esto no me es muy útil, ya que mi análisis se concentrará en los juicios de primer orden, en los que estas cuestiones no surgen.

2. Para un rico análisis de la fenomenología asociada con el pensamiento ocurrente, véase Siewert (1994).

3. Compárese también la observación de Nagel (1974) de que "las características estructurales de la percepción podrían ser más accesibles a la descripción objetiva, aun cuando algo quedaría afuera".

4. Esto está estrechamente relacionado con la hipótesis de Jackendoff de suficiencia computacional: "Toda distinción fenomenológica es causada por/sustentada por/proyectada de una distinción computacional correspondiente" (Jackendoff, 1987, p. 24).

5. Para análisis relacionados con la ceguera visual, véase Tye (1993), Block (1995) y especialmente Dennett (1991).

6. Para un análisis de los peligros de fundir la experiencia consciente y la conciencia *de* una experiencia, y también para una excelente crítica de los enfoques del pensamiento de orden superior en general, véase Siewert (1994). Para un tipo similar de crítica desde un punto de vista reductivo, véase también Dretske (1995).

7. La distinción entre las concepciones de primer orden y de segundo orden refleja la distinción de Nelkin entre los dos conceptos funcionales de la conciencia, C1 y C2 (Nelkin, 1989).

8. Carruthers (1992) examina una propuesta de este tipo. Argumenta

que la *disponibilidad para el pensamiento reflexivo* es naturalmente necesaria y suficiente para una experiencia cualitativa (Carruthers parece tener en mente una variedad más fuerte de disponibilidad, sin embargo, ya que el conductor de camiones desatento de Armstrong no satisface su criterio). En la medida que afirma explícitamente que la conexión sólo es válida con necesidad natural, la propuesta parece no reductiva, aunque Carruthers también caracteriza el enfoque como fisicalista. Alvin Goldman sugirió una concepción similar en conversaciones, cuya intención es ser una caracterización de los estados que son conscientes en sistemas familiares, en lugar de una propuesta reductiva.

9. La distinción entre los registros de primer orden y los juicios de primer orden es paralela a la distinción de Dretske (1995) entre las variedades *fenoménica* y *doxástica* de los estados cognitivos. Esta última corresponde al modo como el sistema considera que las cosas son y la primera corresponde al modo como las cosas están representadas para el sistema. Por supuesto existe una diferencia significativa entre el marco conceptual de Dretske y el mío, en el sentido de que Dretske es esencialmente un funcionalista reductivo (en realidad un teleofuncionalista reductivo) e *identifica* las experiencias con registros de primer orden definidos (teleo)funcionalmente. Me resisto a esa identificación por las razones usuales, pero sigue siendo plausible que las experiencias *correspondan* a registros de primer orden. Mi marco conceptual simplemente posee una distinción extra, reconoce tres tipos diferentes de estados —juicios, registros y estados fenoménicos, donde los registros y los estados fenoménicos están correlacionados pero son distintos— mientras que el marco de Dretske reconoce dos tipos: juicios y estados fenoménicos, es decir, no hace una distinción conceptual entre los estados fenoménicos y los registros correspondientes.

10. Hay un número enorme de cuestiones interesantes acerca del *tipo* de contenido representacional de los registros de primer orden que constituyen la percatación y acerca del tipo de contenido que poseen en paralelo las experiencias correspondientes. Las cuestiones acerca del contenido no son fundamentales para mi análisis, de modo que sólo las planteo aquí brevemente, pero están entre las cuestiones más profundas y sutiles acerca de la experiencia y merecen un tratamiento mucho más detallado en otro lugar.

Una característica central del contenido de la percatación y de la experiencia es que el contenido es aquí, por lo general, *no conceptual*; es contenido que no requiere que un agente posea los conceptos que podrían estar involucrados en la caracterización de ese contenido. Por ejemplo, es plausible que un sistema simple —quizás un perro o un ratón— pueda tener experiencias de color de grano fino, con una representación correspondiente de grano fino de las distinciones de color en el sistema cognitivo, aunque sólo tenga un sistema muy simple de conceptos de color. De un modo similar, en los seres humanos es común que los estados de conciencia y de percatación en la percepción musical tengan contenidos que superan a los conceptos musicales en el repertorio del sujeto.

(Para un análisis del contenido no conceptual, véase Crane [1992], Cussins [1990], Evans [1982], Peacocke [1992]. Parece haber consenso en la

literatura acerca de que el contenido de la experiencia es no conceptual. Una excepción es McDowell [1994], quien utiliza como argumento nuestra capacidad para reidentificar experiencias bajo conceptos como "ese matiz" para llegar a la conclusión de que todo contenido experiencial es conceptual. No es claro que una destreza de este tipo sea un requerimiento para la posesión de la experiencia: es plausible, por ejemplo, que ciertos aspectos sutiles de la experiencia musical en algunos sujetos [por ejemplo, cambios sutiles de clave] pudiesen directamente resistir una conceptualización y reidentificación. Las experiencias en los animales proporcionan otro ejemplo. McDowell parece contento de aceptar la conclusión de que los animales no tienen experiencias, pero podemos encontrar el *modus tollens* al menos tan convincente como el *modus ponens*. Aun si aceptásemos el punto de McDowell, creo que podría rehabilitarse algo semejante a la distinción en cuestión en la forma de dos grados de contenido conceptual.)

Por supuesto, puede haber relaciones *causales* entre los conceptos y la conciencia; no es raro que el cambio conceptual afecte significativamente el carácter de la experiencia. Pero esos recursos conceptuales no parecen ser un *requerimiento* para la experiencia consciente. Lo mismo ocurre con la percatación, en tanto es paralela a la conciencia. El contenido representado por los registros de primer orden que corresponden a las experiencias conscientes, en la percepción visual, por ejemplo, no requiere recursos conceptuales correspondientemente ricos. De este modo, el contenido de la experiencia y de la percatación es en general más primitivo que el de los juicios, cuyos contenidos pueden considerarse más naturalmente como conceptuales.

Una de las cuestiones más interesantes acerca del contenido es si la experiencia tiene contenido representacional de modo intrínseco o si su contenido se deriva de algún modo de un estado cognitivo subyacente. Esta última posición podría ser tentadora, pero no parece del todo correcta: por ejemplo, en este momento parece que mi experiencia visual representa el mundo como algo que mantiene un gran objeto cuadrado frente a mí, y lo hace simplemente en virtud de ser la experiencia que es. Aun una hipotética mente incorpórea que estuviese teniendo una experiencia similar tendría un tipo similar de contenido representacional. Siewert (1994) argumenta de un modo convincente que la experiencia es, por naturaleza, informativa acerca del estado del mundo: una experiencia visual, por ejemplo, es algo que puede evaluarse en cuanto a su precisión (puede representar el mundo correcta o incorrectamente), y también es evaluable en virtud de su propia naturaleza como experiencia visual. De modo que podría ser razonable decir que la experiencia está intrínsecamente cargada de contenido representacional.

Podríamos sentirnos tentados a adoptar un punto de vista inverso, y sostener que el único contenido verdadero está presente en la experiencia, y que el contenido de un registro de primer orden subyacente es, en sí mismo, dependiente del contenido de la experiencia asociada. Podría haber algo en esto, pero tampoco es totalmente satisfactorio; existe un sentido en el cual queremos decir que incluso los registros de primer orden de un zombi representan el mundo como siendo de un determinado modo. Ciertamente

tenemos estados de contenido que no están asociados con experiencias y es difícil determinar si todos nuestros contenidos son de algún modo dependientes de los contenidos de la experiencia. Una posición intermedia que podríamos adoptar es que 1) en cierto sentido el tipo *original* de contenido es el que se encuentra dentro de la experiencia, pero 2) desarrollamos un marco para atribuir contenido a estados cognitivos basándonos, en parte, en la coherencia con el contenido de las experiencias asociadas, y 3) una vez en su lugar, este marco se vuelve autónomo, de modo que podemos hablar del contenido de los estados cognitivos aun en ausencia de experiencias. Esto significaría que las experiencias y los registros asociados podrían, ambos, tener contenido de forma autónoma, sin que exista una extraña y accidental sobredeterminación por la cual el mismo contenido se constituye dos veces. Las cuestiones aquí son bastante sutiles y probablemente merezcan un análisis detallado.

Otra cuestión interesante es si el tipo relevante de contenido es "amplio" (dependiente de objetos en el ambiente) o "estrecho" (dependiente sólo de procesos internos). En la medida en que la experiencia está intrínsecamente cargada de contenido y en la medida en que es superveniente a la organización de un sujeto, el tipo relevante de contenido debe aquí ser estrecho. (Las experiencias podrían tener todavía contenido amplio, pero este no podría ser contenido fijado por la experiencia solamente.) A veces se creyó que el único verdadero contenido representacional es el contenido amplio, pero creo que hay un modo natural de comprender el contenido representacional estrecho (véase Chalmers, 1994c). Podría elaborarse una concepción de este tipo para producir una concepción del contenido estrecho no conceptual de la experiencia y de la percatación como un tipo de contenido que plantea restricciones sobre los mundos centrados que son candidatos a ser el mundo real de un sujeto.

Otra cuestión es si *todas* las experiencias tienen contenido representacional. Es plausible que muchas o la mayoría lo tengan; la mayoría de las experiencias perceptuales parecen ser intrínsecamente informativas acerca del mundo. Existen algunos casos difíciles, sin embargo. ¿Qué hay de los orgasmos, las náuseas o ciertas experiencias emotivas (véase Block [1995] y Tye [1995])? Pero aun en estos casos podríamos encontrar *algún* contenido representacional, ya que las experiencias frecuentemente poseen contenido concerniente a la localización (acá adentro, allá abajo) o cualidad (bueno, malo). No es claro que pueda haber experiencias que están totalmente faltas de contenido representacional; por otro lado, no es obvio que no puedan existir.

Algunos filósofos formularon la propuesta de que las propiedades fenoménicas son sólo propiedades representacionales, de modo que las experiencias se *agotan* en su contenido representacional (por ejemplo, Dretske, 1995; Harman, 1990; Lycan, 1996; Tye, 1992). Frecuentemente esto se formula junto con un enfoque reductivo del contenido representacional, de modo que este enfoque se reduce a una versión del funcionalismo reductivo y resulta poco plausible por las razones usuales. Otra versión de la propuesta podría colocarlo junto con un enfoque no reductivo del contenido

representacional, quizás uno en el cual la única verdadera representación esté en la experiencia. Esto sería más compatible con tomar en serio la conciencia, pero todavía tendría dificultades. En particular, parecería que el contenido representacional podría mantenerse constante entre isomorfos funcionales con inversión de espectro, en cuyo caso la fenomenología sobrepasaría el contenido representacional. Los casos en el párrafo previo también tienden a sugerir que aun en el caso de que todas las experiencias tengan contenido representacional, también tienen características que sobrepasan ese contenido. De modo que no es claro que aun la versión no reductiva de esta propuesta pueda ser exitosa.

11. En comunicación personal y trabajos de próxima publicación.

12 Ocasionalmente, principios como los que mencioné fueron formulados en forma explícita como parte de la metodología del trabajo empírico sobre la mente. No es sorprendente que esto haya ocurrido con mucha frecuencia en el área de la corriente principal de la psicología, que se ocupa principalmente de la experiencia consciente, esto es la *psicofísica*. Por lo general se interpreta que este campo relaciona las propiedades de nuestras sensaciones con las propiedades de los estímulos físicos asociados. Entre los resultados típicos encontramos la ley de Weber-Fechner y la ley potencial de Stevens (Stevens, 1975); estas ofrecen dos modos de relacionar la intensidad de un estímulo con la intensidad de la sensación correspondiente. Aunque a veces se sostiene que los explananda primarios en la psicofísica son los datos de tercera persona, como los informes subjetivos, es innegable que características de la experiencia de primera persona - como la experiencia de ciertas ilusiones ópticas - están entre los fenómenos centrales que el campo intenta explicar.

(Horst [1995] realiza una fuerte defensa de que los datos primarios en este campo son frecuentemente experiencias de primera persona de diversos fenómenos como las ilusiones. En conferencias, por ejemplo, los investigadores le asignan mucha importancia a poder "ver" por sí mismos diversos efectos. Podríamos también argumentar que las disputas entre los enfoques de Fechner y Stevens sobre la medición de la sensación [véase Stevens, 1975] sólo tienen sentido si suponemos que existe un objetivo de medir un objeto común, la experiencia fenoménica; si esto no fuera así, simplemente tendríamos mediciones no competidoras de diferentes fenómenos funcionales.)

Dentro de la psicofísica, existieron discusiones ocasionales sobre los modos mediante los cuales las observaciones empíricas pueden ayudar en la explicación de las sensaciones subjetivas. Algunos investigadores han sido llevados a formalizar principios explícitos sobre los cuales este trabajo se apoya, conocidos diversamente como "hipótesis psicofísicas de enlace" (Brindley, 1960) o "proposiciones generales de enlace" (Teller, 1984). Los "axiomas de la correspondencia psicofísica" formulados por Müller (1896; citado en Boring, 1942, p. 89) son un buen ejemplo:

1. La base de todo estado de conciencia es un proceso material, un denominado proceso psicofísico, a cuya ocurrencia se reúne la presencia del estado consciente.

2. A una igualdad, similitud o diferencia en la constitución de las sensaciones... le corresponde una igualdad, similitud o diferencia en la constitución del proceso psicofísico y recíprocamente. Más aún, a una mayor o menor similitud de sensaciones, también le corresponde respectivamente una mayor o menor similitud del proceso psicofísico y recíprocamente.

3. Si los cambios que una sensación atraviesa tienen la misma dirección, o si las diferencias que existen entre series de sensaciones son de una dirección similar, entonces los cambios que atraviesa el proceso psicofísico, o las diferencias del proceso psicofísico dado, tienen una dirección similar. Más aún, si una sensación es variable en n direcciones, entonces el proceso psicofísico subyacente también debe ser variable en n direcciones y recíprocamente.

Es claro que estos principios están estrechamente relacionados con el principio de coherencia estructural. A pesar de ciertas diferencias en el lenguaje, podemos ver que todas las proposiciones similares a las de arriba son consecuencias directas del principio de coherencia; y, juntas, constituyen buena parte de su fuerza. De modo que una vez más podemos ver que el principio de coherencia estructural y sus variantes desempeñan un papel central en hacer posible que la investigación empírica produzca concepciones explicativas de diversas características de la experiencia.

No es sorprendente que la naturaleza de estos principios haya sido debatida dentro de la psicofísica de un modo paralelo a los tipos de controversias que se encuentran en la filosofía de la mente (véase por ejemplo, Brindley, 1960; Marks, 1978; D. Teller, 1984, 1990). Algunos las consideraron hipótesis empíricas, pero no parece que puedan derivarse o falsificarse por medio de pruebas empíricas, al menos de la variedad de tercera persona. Otros, especialmente los de tendencia operacionalista los consideraron aseveraciones definicionales; esto corresponde a una posición funcionalista reductiva en la filosofía. Con frecuencia, simplemente se los tomó como supuestos generales, o premisas, concernientes a la naturaleza de la conexión psicofísica. En cualquier caso, la ciencia se las ha arreglado para proceder exitosamente sin ninguna resolución real de estas cuestiones. Para propósitos explicativos, la forma del puente es más importante que su condición metafísica.

En general, las cuestiones "filosóficas" acerca de la relación entre los procesos físicos y la experiencia burbujean justo por debajo de la superficie en muchas discusiones teóricas en la psicofísica. Por lo que yo sé, esto no encontró mucha discusión en la literatura filosófica (aunque véase Savage [1970] para una crítica filosófica de la metodología de la medición de la sensación). Es probable que constituya un importante tema para un estudio extenso.

## Capítulo 7

1. Lycan (1987), por ejemplo.

2. Churchland y Churchland (1981) objetaron los argumentos de la "nación china" sobre la base de que un sistema de este tipo necesitaría

manejar alrededor de $10^{30.000.000}$ entradas a la retina y un número aun más vasto de estados internos del cerebro. La simulación de la población, que requiere una persona por entrada y una persona por estado, requeriría por lo tanto una cantidad de personas mucho mayor de lo que una población podría proporcionar.

Esta objeción pasa por alto el hecho de que las entradas y los estados internos están estructurados en forma combinatoria. En lugar de representar cada patrón de entrada (en más de $10^8$ células) con una sola persona, lo que requeriría $2^{10^8}$ personas, sólo necesitamos $10^8$ personas para representar la entrada de un patrón estructurado. Lo mismo es válido para los estados internos. Por lo tanto no necesitamos más personas que la cantidad de células en el cerebro.

3. Bogen (1981) y Lycan (1987) hacen la sugerencia de que estas situaciones "accidentales" no tendrían qualia, ya que los qualia requieren teleología. Esto tendría la extraña consecuencia de hacer que la presencia o ausencia de los qualia dependa de la historia de un sistema. Creo que sería mejor aceptar que un sistema de este tipo tendría qualia, al mismo tiempo que señalamos la improbabilidad de que un sistema de esta clase pueda surgir por azar.

4. Analizo esta cuestión más detalladamente en Chalmers (1994a).

5. Esta cifra surge de advertir que hay $10^{10^9}$ posibles elecciones para el consecuente de cada condicional, que representa el estado global al que pasará el sistema. La probabilidad de que un estado global dado pase al estado correcto siguiente es entonces de 1 en $10^{10^9}$. De hecho será menor, ya que cualquier estado global determinado será realizable por muchos estados "maximales" diferentes del sistema físico, cada uno de los cuales se requiere que tenga una transición apropiada. Hay $10^{10^9}$ condicionales de este tipo que deben satisfacerse, de donde surge la cifra de más arriba.

6. Para alguna fábulas relacionadas, véase también Harrison (1981a, 1981b).

7. Quizás algún elemento de esta situación pueda explicarse mediante la constitución de creencias por la experiencia del tipo que analizamos en el apartado 7 del capítulo 5: tal vez el concepto de "experiencia de rojo" de José ahora refiera a experiencias de rosa, por ejemplo, de modo que él estaría totalmente equivocado. Esta estrategia ciertamente no ayudaría con los errores en sus juicios acerca de las distinciones, sin embargo.

8. Cuda (1985) afirma que una descripción de sistemas con creencias equivocadas de esta clase *no tiene sentido.* No ofrece ningún argumento para ello aparte de la afirmación de que si la descripción tuviese sentido, entonces tendría sentido pensar que *nosotros* estamos equivocados de un modo semejante, lo que (él dice) claramente no es así. Pero esto me parece un argumento falaz. Tiene *sentido* suponer que yo me podría equivocar de ese modo y que la hipótesis es coherente; sólo sucede que mi situación epistémica me muestra que la hipótesis no es *verdadera* en mi propio caso, porque tengo experiencia directa de los qualia de rojo brillante y similares.

9. Esto no implica que no puedan hallarse argumentos sorites en este campo; por ejemplo, véase Tienson (1987).

10. Esta posición está muy estrechamente asociada con Shoemaker (1982), pero también fue defendida por Horgan (1984a), Putnam (1981) y otros más.

11. Podría incluso ocurrir que estos casos existan en el mundo real. En un artículo interesante, Nida-Rümelin (1996) hace notar que la investigación en la visión del color nos lleva a esperar que debería de haber casos de visión en color "pseudonormal", en los que 1) los conos R en la retina tengan el patrón de respuesta usualmente asociado a los conos G y 2) los conos G tengan el patrón de respuesta usualmente asociado a los conos R. Tomados separadamente, 1) y 2) son las causas estándar de la ceguera al color rojo-verde. En teoría, las anomalías genéticas responsables de 1) y 2) podrían ocurrir juntas, lo que produciría un sujeto que se comportará en una forma conductualmente muy similar a una persona normal, pero que puede tener experiencias de color que estén invertidas respecto del resto de la población.

12. Putnam (1981) y Shoemaker (1982) utilizan este ejemplo para oponerse a las concepciones funcionalistas de los qualia pero, en el mejor de los casos, sus argumentos contradicen un principio de invariancia "de grano grueso", en el cual, por ejemplo, se sostiene que el mismo tipo de experiencia surge siempre de estados que son desencadenados por cosas azules y llevan a informes de "azul". El principio de grano fino no está amenazado (Levine [1988] ofrece un argumento relacionado).

Existe una variante de este escenario en el cual los sujetos reconectados sufren un proceso de adaptación, aprendizaje y finalmente amnesia (olvidan que las cosas alguna vez parecieron diferentes), y terminan en un estado conductualmente idéntico al original. Sin embargo, hay pocas razones para creer que serán organizacionalmente idénticos, en especial dado que la reconexión todavía se refleja en el estado de su cerebro. Si por algún proceso especial la organización termina exactamente como comenzó, no parecería inverosímil que las experiencias deban también revertir a su estado original (Cole [1990] y Rey [1992] defienden versiones de una hipótesis de reversión aquí).

13. Un argumento relacionado fue formulado por Seager (1991, pp. 39-41), quien describe un caso en el cual las células retinianas están "ajustadas" a un nivel más alto en el espectro óptico. Este argumento puede tratarse del mismo modo que el de Block.

14. Block responde a una objeción relacionada haciendo notar que podemos mover la "lente" hacia adentro en el sistema, reconectando cosas en el nervio óptico o en la corteza visual, por ejemplo. Como antes, sin embargo, esto no representa ningún caso de isomorfos organizacionales con experiencias diferentes. La descripción de Block de esos casos parece directamente compatible con el enfoque de que los qualia dependen de la organización de los sistemas "centrales". Se nos pide que creamos que las experiencias son las mismas en ciertos casos precisamente porque el procesamiento central no se encuentra afectado; se supone que nosotros creemos que las experiencias difieren en los casos en los que el procesamiento central difiere. Los argumentos de esta clase no pueden refutar el principio de invariancia.

15. El argumento en este apartado está lejanamente inspirado en la

historia de Dennett "¿Dónde estoy yo?" (1978d). Shoemaker (1982) considera una situación con una cierta semejanza a la que describo más abajo. Puede encontrarse un análisis más estrechamente relacionado en Seager (1991, p. 43), aunque este autor no defiende el principio de invariancia. Cuando este libro entraba en imprenta, descubrí un estimulante y reciente artículo de Arnold Zuboff (1994) que formula lo que es esencialmente el argumento de los qualia danzantes para apoyar una versión del funcionalismo reductivo, en donde sostiene que los qualia danzantes son imposibles *a priori*.

16. White (1986) sostiene algo similar, sugiriendo que si diferencias físicas no funcionales son significativas para los qualia, entonces aun diminutas diferencias en el ADN podrían también afectarlos.

17. Shoemaker (1982) enuncia un criterio complejo de lo específica que debe ser una propiedad fisiológica para fijar los qualia o para "realizar una quale", como dice. Sin embargo, me parece que si mi análisis aquí ha sido correcto, su criterio seleccionará una propiedad funcional de grano fino.

18. Esta sugerencia fue hecha por Terry Horgan en conversación personal.

## Capítulo 8

1. Otros que sugirieron vínculos entre la conciencia y la información fueron Bohm (1980), Sayre (1976) y Velmans (1991), aunque los detalles de sus propuestas son bastante diferentes de las mías. La idea de Sayre de un "monismo neutral" de información es bastante sugerente, sin embargo (agradezco a Steve Horst por señalármelo). Algo bastante parecido al principio del doble aspecto se analiza en Lockwood (1989), cap. 11, aunque él no lo formula en términos de información.

2. Un trabajo no publicado mío (Chalmers, 1990) se ocupa de esta estrategia en la comprensión de la relación entre la conciencia y los juicios acerca de la misma, y la utiliza para formular una "teoría" básica de la conciencia (que involucra patrones de información) que es una predecesora de algunas de las ideas de este capítulo. En ese trabajo, denomino al requerimiento de coherencia explicativa el "Test de coherencia", que cualquier teoría de la conciencia debe pasar.

3. Sin embargo, para la tesis de que los termostatos tienen creencias y deseos, véase McCarthy (1979).

4. El enfoque de Wheeler se concentra en resultados de medición o "respuestas a preguntas sí/no" como la base de todo y, como tal, puede estar más cerca de una forma de idealismo que el enfoque que yo formulo aquí.

5. Véase también la interesante exposición de las ideas de Fredkin en Wright (1988), para más detalles acerca de la metafísica subyacente.

6. Existen estimulantes exposiciones de estos problemas para el enfoque russelliano en Foster (1991, pp. 119-30) y Lockwood (1992).

7. Lockwood (1992) sugiere que un enfoque russelliano puede invocar leyes primitivas con esta finalidad. No enfrenta la objeción de que la introducción de nuevas leyes con esta función parece comprometer los

atractivos originales del enfoque russelliano. Por ejemplo, esta introducción plantea problemas de epifenomenalismo que el enfoque russelliano prometía evitar, y también requiere una considerable expansión de la ontología más allá de las propiedades intrínsecas necesarias para fundamentar la física. (Debería notarse que Lockwood no se basa en esas leyes para resolver el problema de granularidad; su idea principal acerca de resolver el problema es una interesante sugerencia que involucra a la mecánica cuántica.)

## Capítulo 9

1. El material en este apartado ha sido extraído fundamentalmente de Chalmers (1994a).

2. Putnam (1988, pp. 120-25) ofrece un argumento separado en favor de la conclusión de que cualquier sistema abierto ordinario implementa cualquier autómata de estado finito. Analizo en detalle este argumento en Chalmers (1995a). Cuando se lo examina, el argumento parece obtener su fuerza de permitir que los condicionales físicos de transición de estados en la definición de implementación carezcan de fuerza modal.

3. Analizo esta forma de comprender el papel explicativo de la computación en la ciencia cognitiva en Chalmers (1994b).

4. Korb (1991) y Newton (1989) proponen cuestiones relacionadas. Ambos sugieren que el cuarto chino podría proporcionar un buen argumento en contra de la conciencia de las máquinas, si no en contra de su intencionalidad.

5. Hofstadter (1981) formula un espectro similar de casos intermedios entre un cerebro y el cuarto chino.

6. La idea de que el homunculus en el cuarto chino es análoga a un demonio que corre alrededor del cráneo fue sugerida por Haugeland (1980).

7. Es notable que aunque Dreyfus (1972) tituló su libro haciendo referencia a este tipo de objeción, *What Computers Can't Do* [Lo que los ordenadores no pueden hacer], luego acepta que el *tipo* correcto de sistema computacional (por ejemplo, un sistema conexionista) escaparía a esas objeciones. En efecto, "lo que los ordenadores pueden hacer" se identifica con lo que una clase muy estrecha de sistemas computacionales puede hacer.

8. Esta objeción directa a los argumentos gödelianos fue publicada por primera vez en Putnam (1960), creo, y por lo que yo sé nunca ha sido refutada a pesar de los mejores esfuerzos de Lucas y Penrose. Penrose (1994, sec. 3.3) argumenta que debe poder determinar la consistencia del sistema formal que captura su propio razonamiento, ya que puede seguramente determinar la verdad de los axiomas y la validez de las reglas de inferencia. Esto parece depender del supuesto de que el sistema computacional es un sistema de axiomas más reglas en primer lugar, lo que no es necesario en el caso general (considérese la simulación neuronal del cerebro). Aun en el caso de axiomas más reglas, no me resulta claro de que podamos determinar la validez de *toda* regla que nuestro sistema podría utilizar, especialmente de aquellas que se aplican a los límites externos de la enumeración ordinal en la gödelización

iterada, que es donde los argumentos gödelianos en el caso humano realmente tendrán su fuerza.

9. Probablemente sea una buena idea hacer esto, en caso de que un patrón específico de redondeos en el nivel de $10^{-10}$ produzca una distribución sesgada de la conducta. Para estar seguros, dado que existe ruido en el nivel $10^{-10}$, podríamos aproximar el sistema en el nivel $10^{-20}$, aproximando la distribución del ruido también en ese nivel.

10. A veces —por lo general sólo en la filosofía de la mente— términos como "computación" se utilizan para referir exclusivamente a la clase de computaciones simbólicas o a las computaciones sobre representaciones (esto es, sistemas en los cuales los objetos sintácticos básicos son también objetos semánticos básicos). Por supuesto, poco depende de esta cuestión terminológica: lo importante desde el punto de vista de la inteligencia artificial es que haya algún tipo de sistema formal tal que la implementación sea suficiente para la mentalidad, se la considere o no una "computación" según ese criterio. Sin embargo, debe notarse que, en cualquier caso, utilizar el término de este modo es perder contacto con sus orígenes en la teoría de la computación. Incluso la mayoría de las máquinas de Turing no serán "computacionales" en ese sentido, ya que sólo unas pocas de estas puede interpretarse que realizan computaciones sobre representaciones conceptuales. Por razones similares, limitar la clase de "computaciones" de esa manera es perder la *universalidad* (Church-Turing) de la computación, la que constituye, tal vez, la mejor razón para creer en la tesis de la IA (funcional) en primer lugar.

## Capítulo 10

1. Al menos para una partícula de spin $\frac{1}{2}$ como el electrón. Dejo de lado los casos en los que el spin tiene otros valores básicos.

2. Simplifico aquí, como en otros lugares. Ninguna medición es perfectamente precisa, de manera que nunca puede surgir un estado con una posición verdaderamente definida. En cambio, la función de onda colapsará en un estado en el cual toda la amplitud está concentrada en un intervalo muy pequeño de localizaciones. No obstante, es más fácil hablar como si las posiciones colapsadas estuvieran verdaderamente definidas.

3. Una densidad probabilística en el caso continuo.

4. Albert (1992) sugiere que el término "conciencia" es tan vago como "medición" y "macroscópico"; pero, me parece que este criterio es atractivo en parte porque es plausible que *existan* hechos empíricos acerca de si un sistema es consciente.

5. Me baso en Albert y Loewer (1990) y Albert (1992) para mi exposición de la interpretación GRW.

6. Nótese que este es el único punto en el capítulo en el que surgen el teorema de Bell y los resultados de Einstein-Podolsky-Rosen (EPR). A veces se considera que estos resultados son la fuente principal de los problemas filosóficos de la mecánica cuántica, pero yo creo que los problemas surgen previamente a las consideraciones EPR. Aun sin EPR, tendríamos la difícil

elección entre el colapso, las variables ocultas y Everett. EPR simplemente incrementan las dificultades de las teorías de variables ocultas, al mostrar que ellas (como las del colapso) deben ser no locales; posiblemente también incremente el atractivo de la interpretación de Everett, que es la única interpretación local compatible con el resultado.

7. El enfoque del único gran mundo parece ser la interpretación más común de la perspectiva de Everett entre los físicos (en especial entre los cosmólogos cuánticos, que utilizan este marco conceptual todo el tiempo). La interpretación del "desdoblamiento de mundos" es principalmente un artefacto de las popularizaciones. A veces incluso los defensores del enfoque de único gran mundo hablan de "desdoblamiento", pero este es sólo un modo vívido de hablar acerca del hecho de que una función de onda evoluciona en una superposición. No existe ningún proceso especial de desdoblamiento de mundos; cuanto más, existe una especie de división local de la función de onda. En cualquier caso creo que es mejor evitar el discurso sobre "desdoblamiento", ya que inevitablemente promueve la confusión.

8. Esta objeción fue planteada por Bell (1981), Bohm y Hiley (1993) y Hodgson (1988), entre muchos otros.

9. La estrategia más razonable podría ser realizar mi apuesta de acuerdo con un dispositivo cuántico que produce una respuesta "no" con probabilidad 0,999 y "sí" con probabilidad 0,001. De este modo, si el enfoque de Everett es falso casi seguramente estaré en lo correcto y si es verdadero, al menos *una* de mis mentes descendientes sobrevivirá.

# Bibliografía


Ackerman, D. *A Natural History of the Senses*, Nueva York, Random House, 1990. [*Una historia natural de los sentidos*, Barcelona, Anagrama, 1992.]

Adams, R. M. "Theories of actuality", *Nous*, *8*, 1974, 211-31.

Akins, K. "What it is like to be boring and myopic?", en B. Dahlbom (comp.), *Dennett and His Critics*, Oxford, Blackwell, 1993.

Albert, D. *Quantum Mechanics and Experience*, Cambridge, Mass., Harvard University Press, 1992.

Albert, D. y B. Loewer. "Interpreting the many-worlds interpretation", *Synthese*, *77*, 1988, 195-213.

Albert, D. "Two no-collapse interpretations of quantum mechanics", *Nous*, 23, 1989, 169-86.

Albert, D. "Wanted dead or alive: Two attempts to solve Schrödinger's paradox", *PSA 1990*, vol. 1, 1990, 277-85.

Alexander, S. *Space, Time, and Deity*, Londres, Macmillan, 1920.

Armstrong, D. M. *A Materialist Theory of the Mind*, Londres, Routledge and Kegan Paul, 1968.

Armstrong, D. M. *Belief, Truth, and Knowledge,* Cambridge, Cambridge University Press, 1973.

Armstrong, D. M. "What is consciousness?", en *The Nature of Mind*, Ithaca, N.Y., Cornell University Press, 1981.

Armstrong, D. M. "Metaphysics and supervenience", *Critica*, *42*, 1982, 3-17.

Armstrong, D. M. *What Is a Law of Nature?*, Cambridge, Cambridge University Press, 1983.

Armstrong, D. M. *A Combinatorial Theory of Possibility*, Cambridge, Cambridge University Press, 1990.

Austin, D. F. *What's the Meaning of "This"?*, Ithaca, N.Y., Cornell University Press, 1990.

Baars, B. J. *A Cognitive Theory of Consciousness*, Cambridge, Cambridge University Press, 1988.

Bacon, J. "Supervenience, necessary coextension, and reducibility", *Philosophical Studies*, *49*, 1986, 163-176.

Barwise, J. y J. Perry. *Situations and Attitudes*, Cambridge, Mass., MIT Press, 1983. [*Situaciones y actitudes.* Madrid, Visor, 1992.]

Bateson, G. *Steps to an Ecology of Mind*, San Francisco, Chandler, 1972. [*Pasos hacia una ecología de la mente - Una aproximación revolucionaria a la autocomprensión del hombre.* Buenos Aires, Ediciones Carlos Lohlé, 1976.]

Bealer, G. "Mental properties", *Journal of Philosophy*, *91*, 1994, 185-208.

Bell, J. S. "On the Einstein-Podolsky-Rosen paradox", *Physics*, *1,* 1964, 195-200. [Reimpreso en Bell, 1987b]

Bell, J. S. "The measurement theory of Everett and de Broglie's pilot wave", en M. Flato (comp.), *Quantum Mechanics, Determinism, Causality, and Particles*, Dordrecht, Reidel, 1976. [Reimpreso en Bell, 1987b]

Bell, J. S. "Quantum mechanics for cosmologists", en C. Isham, R. Penrose y D. Sciama (comps.), *Quantum Gravity*, Vol. 2, Oxford, Oxford University Press, 1981. [Reimpreso en Bell, 1987b.]

Bell, J. S. "Are there quantum jumps?", en *Schrödinger: Centenary of a Polymath*, Cambridge, Cambridge University Press, 1987a.

Bell, J. S. *Speakable and Unspeakable in Quantum Mechanics*, Cambridge, Cambridge University Press, 1987b. [*Lo decible y lo indecible en mecánica cuántica.* Madrid, Alianza, 1990.]

Bigelow, J., y R. Pargetter. "Acquaintance with qualia", *Theoria*, *56*, 1990, 129-47.

Bisiach, E. "The (haunted) brain and consciousness, en A. Marcel y E. Bisiach (comps.), *Consciousness in Contemporary Science*, Oxford, Oxford University Press, 1988.

Blackburn, S. "Moral realism", en J. Casey (comp.), *Morality and Moral Reasoning*, Londres, Methuen, 1971.

Blackburn, S. "Supervenience revisited", en I. Hacking (comp.), *Exercises in Analysis: Essays by Students of Casimir Lewy*, Cambridge, Cambridge University Press, 1985.

Blackburn, S. "Filling in space", *Analysis*, *50*, 1990, 62-65.

Block, N. "Troubles with functionalism", en C.W. Savage (comp.), *Perception and Cognition: Issues in the Foundation of Psychology*, Minneapolis, University of Minnesota Press, 1978. [Reimpreso en N. Block (comp.), *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge, Mass., Harvard University Press, 1980.]

Block, N. "What is functionalism?", en N. Block (comp.), *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge, Mass., Harvard University Press, 1980.

Block, N. "Psychologism and behaviorism", *Philosophical Review*, *90*, 1981, 5-43.

Block, N. "Inverted earth", *Philosophical Perspectives*, *4*, 1990, 53-79.

Block, N. "On a confusion about a function of consciousness", *Behavioral and Brain Sciences*, *18*, 1995, 227-47.

Boden, M. "Escaping from the Chinese Room", en *Computer Models of Mind*, Cambridge, Cambridge University Press, 1988.

Bogen, J. "Agony in the schools", *Canadian Journal of Philosophy*, *11*, 1981, 1-21.

Bohm, D. "A suggested interpretation of quantum mechanics in terms of 'hidden variables'", partes 1 y 2, *Physical Review*, *85*, 1952, 166-193.

Bohm, D. *Wholeness and the Implicate Order*, Londres, Routledge, 1980. [*La totalidad y el orden implicado*. Barcelona, Kairós, 3ª ed., 1998.]
Bohm, D. y B. Hiley. *The Undivided Universe: An Ontological Interpretation of Quantum Theory*, Londres, Routledge, 1993.
Boring, E. G. *Sensation and Perception in the History of Experimental Psychology*, Nueva York, Appleton-Century-Crofts, 1942.
Boyd, R. N. "Materialism without reductionism: What physicalism does not entail, en N. Block (comp.), *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge, Mass., Harvard University Press, 1980.
Boyd, R. N. "How to be a moral realist", en G. Sayre-McCord (comp.), *Essays on Moral Realism*, Ithaca, N.Y., Cornell University Press, 1988.
Brindley, G. S. *Physiology of the Retina and Visual Pathway*, Londres, Edward Arnold, 1960.
Brink, D. *Moral Realism and the Foundations of Ethics*, Cambridge, Cambridge University Press, 1989.
Broad, C. D. *Mind and Its Place in Nature*, Londres, Routledge and Kegan Paul, 1925.
Brooks, D. H. M. "How to perform a reduction", *Philosophy and Phenomenological Research, 54*, 1994, 803-14.
Byrne, A. *The emergent mind*, Tesis de doctorado, Princeton University, 1993.
Campbell, K. K. *Body and Mind*, Nueva York, Doubleday, 1970.
Carroll, J. W. "The Humean tradition", *Philosophical Review, 99*, 1990, 185-219.
Carroll, J. W. *Laws of Nature*, Cambridge, Cambridge University Press, 1994.
Carruthers, P. "Consciousness and concepts", *Proceedings of the Aristotelian Society*, supl., *66,* 1992, 41-59.
Chalmers, D. J. "Consciousness and cognition", Technical Report 38, Center for Research on Concepts and Cognition, Indiana University, 1990.
Chalmers, D. J. "On implementing a computation", *Minds and Machines, 4*, 1994a, 391-402.
Chalmers, D. J. "A computational foundation for the study of cognition", PNP Technical Report 94-03, Washington University, 1994b.
Chalmers, D. J. "The components of content", PNP Technical Report 94-04, Washington University, 1994c. [http://www.artsci.wustl.edu/nphilos/pnp.html]
Chalmers, D. J. "Does a rock implement every finite state automaton?", *Synthese*, 1995a.
Chalmers, D. J. "Facing up to the problem of consciousness", *Journal of Consciousness Studies, 2*, 1995b, 200-219. [También en S. Hameroff, A. Kaszniak y A. Scott (comps.), *Toward a Science of Consciousness*, Cambridge, Mass., MIT Press, 1996.]
Chalmers, D. J. "Minds, machines, and mathematics", *PSYCHE, 2*, 1, 1995c.
Chalmers, D. J. "The puzzle of conscious experience", *Scientific American, 273*, 1995d, 80-86.
Cheney, D. L. y R. M. Seyfarth. *How Monkeys See the World*, Chicago, University of Chicago Press, 1990.
Chisholm, R. *Perceiving*, Ithaca, N.Y., Cornell University Press, 1957.

Churchland, P. M. “Reduction, qualia and the direct introspection of brain states, *Journal of Philosophy, 82*, 1985, 8-28.
Churchland, P.M. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain,* Cambridge, Mass., MIT Press, 1995.
Churchland, P.M. y P. S. Churchland. “Functionalism, qualia and intentionality”, *Philosophical Topics, 12,* 1981, 121-32.
Churchland, P. S. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*, Cambridge, Mass., MIT Press, 1986.
Churchland, P. S. “The significance of neuroscience for philosophy”, *Trends in the Neurosciences, 11*, 1988, 304-7.
Clark, A. “Psychofunctionalism and chauvinism”, *Philosophy of Science, 53,* 1986, 535-59.
Clark, A. *Sensory Qualities*, Oxford, Oxford University Press, 1993.
Cole, D. “Functionalism and inverted spectra”, *Synthese, 82*, 1990, 207-22.
Conee, E. “Physicalism and phenomenal properties”, *Philosophical Quarterly, 35*, 1985a, 296-302.
Conee, E. “The possibility of absent qualia”, *Philosophical Review, 94,* 1985b, 345-66.
Cowey, A. y P. Stoerig. Reflections on blindsight, en D. Milner y M. Rugg (comps.), *The Neuropsychology of Consciousness*, Londres, Academic Press, 1992.
Crane, T. “The nonconceptual content of experience”, en T. Crane (comp.), *The Contents of Experience,* Cambridge, Cambridge University Press, 1992.
Crick, F. H. C. *The Astonishing Hypothesis: The Scientific Search for the Soul*, Nueva York, Scribner, 1994. [*La búsqueda científica del alma: una revolucionaria hipótesis para el siglo XXI*. Barcelona, Círculo de lectores, 1994; *La búsqueda científica del alma.* Madrid, Editorial Debate, 3ª ed., 1994.]
Crick, F. H. C. y C. Koch. “Towards a neurobiological theory of consciousness”, *Seminars in the Neurosciences, 2*, 1990, 263-75.
Cuda, T. “Against neural chauvinism”, *Philosophical Studies, 48,* 1985, 111-27.
Cussins, A. “The connectionist construction of concepts, en M. Boden (comp.), *The Philosophy of Artificial Intelligence*, Oxford, Oxford University Press, 1990.
Daneri, A., A. Loinger y G. M. Prosperi. “Quantum theory of measurement and ergodicity conditions, *Nuclear Physics, 33*, 297-319, 1962. [Reimpreso en Wheeler y Zurek, 1983.]
Davidson, D. “Mental events”, en L. Foster y J. Swanson (comps.), *Experience and Theory*, Londres, Duckworth, 1970.
Davies, M. K. y I. L. Humberstone. “Two notions of necessity”, *Philosophical Studies, 38*, 1980, 1-30.
Dennett, D.C. *Content and Consciousness*, Londres, Routledge and Kegan Paul, 1969. [*Contenido y conciencia.* Barcelona, Gedisa, 1996.]
Dennett, D.C. *Brainstorms*, Cambridge, Mass., MIT Press, 1978a.
Dennett, D.C. “Are dreams experiences?”, en Dennett (1978a), 1978b.
Dennett, D.C. “Toward a cognitive theory of consciousness”, en Dennett (1978a), 1978c.

Dennett, D.C. "Where am I?", en Dennett (1978a), 1978d.
Dennett, D.C. "On the absence of phenomenology", en D. Gustafson y B. Tapscott (comps.), *Body, Mind, and Method*, Dordrecht, Kluwer, 1979.
Dennett, D.C. *The Intentional Stance*, Cambridge, Mass., MIT Press, 1987. [*La actitud intencional*, Barcelona, Gedisa, 1991.]
Dennett, D.C. "Quining qualia", en A. Marcel y E. Bisiach (comps.), *Consciousness in Contemporary Science*. Oxford, Oxford University Press, 1988.
Dennett, D.C. *Consciousness Explained*, Boston, Little, Brown, 1991.
Dennett, D.C. "Back from the drawing board", en B. Dahlbom (comp.), *Dennett and His Critics*, Oxford, Blackwell, 1993a.
Dennett, D.C. "The message is: There is no medium", *Philosophy and Phenomenological Research, 53*, 1993b, 919-31.
Descartes, R. *The Philosophical Writings of Descartes* (Traducido por J. Cottingham, R. Stoothoff y D. Murdoch), Cambridge, Cambridge University Press, 1984.
DeWitt, B. S. "Quantum mechanics and reality", *Physics Today*, *23*, 1970, 30-35. [Reimpreso en DeWitt y Graham, 1973.]
DeWitt, B. S. "The many-universes interpretation of quantum mechanics", en B. d'Espagnat (comp.), *Foundations of Quantum Mechanics*, Nueva York, Academic Press, 1971. [Reimpreso en DeWitt y Graham, 1973.]
DeWitt, B. S. y N. Graham (comps.). *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton, Princeton University Press, 1973.
Dretske, F. I. "Laws of nature", *Philosophy of Science, 44*, 1977, 248-68.
Dretske, F. I. *Knowledge and the Flow of Information*, Cambridge, Mass., MIT Press, 1981. [*Conocimiento y flujo de información*. Barcelona, Salvat, 1989.]
Dretske, F. I. *Naturalizing the Mind*, Cambridge, Mass., MIT Press, 1995.
Dreyfus, H. *What Computers Can't Do*, Nueva York, Harper & Row, 1972.
Eccles, J. C. "Do mental events cause neural events analogously to the probability fields of quantum mechanics?", *Proceedings of the Royal Society of London, B227*, 1986, 411-28.
Edelman, G. *The Remembered Present: A Biological Theory of Consciousness*, Nueva York, Basic Books, 1989.
Edelman, G. *Bright Air, Brilliant Fire*, Nueva York, Basic Books, 1992.
Elitzur, A. "Consciousness and the incompleteness of the physical explanation of behavior", *Journal of Mind and Behavior, 10*, 1989, 1-20.
Evans, G. "Reference and contingency", *The Monist*, 62, 1979, 161-89.
Evans, G. *The Varieties of Reference*, Oxford, Oxford University Press, 1982.
Everett, H. "'Relative-state' formulations of quantum mechanics", *Reviews of Modern Physics, 29*, 1957, 454-62. [Reimpreso en Wheeler y Zurek, 1983.]
Everett, H. "The theory of the universal wave function", en B.S. de Witt y N. Graham (comps.). *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton, Princeton University Press, 1973.
Farah, M. "Visual perception and visual awareness after brain damage: A tutorial overview", en C. Umilta y M. Moscovitch (comps.), *Conscious and Nonconscious Information Processing: Attention and Performance 15*, Cambridge, Mass., MIT Press, 1994.

Farrell, B. A. "Experience, *Mind, 59*, 1950, 170-98.
Feigl, H. "The 'mental' and the 'physical'", en H. Feigl, M. Scriven y G. Maxwell (comps.), *Concepts, Theories, and the Mind-Body Problem*, Minnesota Studies in the Philosophy of Science, vol. 2, Minneapolis, University of Minnesota Press, 1958.
Feldman, F. "Kripke on the identity theory", *Journal of Philosophy, 71*, 1974, 665-76.
Field, H. "Theory change and the indeterminacy of reference", *Journal of Philosophy, 40,* 1973, 762-81.
Flanagan, O. *Consciousness Reconsidered*, Cambridge, Mass., MIT Press, 1992.
Fodor, J. A. "Searle on what only brains can do", *Behavioral and Brain Sciences, 3,* 1980, 431-32.
Fodor, J. A. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, Mass., MIT Press, 1987. [*Psicosemántica.* Madrid, Tecnos, 1994.]
Fodor, J. A. "The big idea: Can there be a science of mind?", *Times Literary Supplement*, julio 3, 1992, pp. 5-7.
Forrest, P. "Ways worlds could be", *Australasian Journal of Philosophy, 64*, 1986, 15-24.
Foss, J. "On the logic of what it is like to be a consciousness subject", *Australasian Journal of Philosophy, 67*, 1989, 305-20.
Foster, J. *The Immaterial Self: A Defence of the Cartesian Dualism Conception of Mind*, Londres, Routledge, 1991.
Fredkin, E. "Digital mechanics", *Physica, D45*, 1990, 254-70.
Geach, P. *Mental Acts*, Londres, Routledge and Kegan Paul, 1957.
Gell-Mann, M. y J. B. Hartle. "Quantum mechanics in the light of quantum cosmology", en W. Zurek (comp.), *Complexity, Entropy, and the Physics of Information,* Redwood City, Calif., Addison-Wesley, 1990.
Gert, B. "Imagination and verifiability", *Philosophical Studies, 16*, 1965, 44-47.
Ghirardi, G. C., A. Rimini y T. Weber. "Unified dynamics for microscopic and macroscopic systems", *Physical Review, D34*, 1986, 470.
Goldman, A. *Epistemology and Cognition*, Cambridge, Mass., Harvard University Press, 1986.
Goldman, A. "The psychology of folk psychology", *Behavioral and Brain Sciences, 16*, 1993, 15-28.
Gunderson, K. "Asymmetries and mind-body perplexities", en M. Radner y S. Winokur (comps.), *Analyses of Theories and Methods of Physics and Psychology*, Minnesota Studies in the Philosophy of Science, vol. 4, Minneapolis, University of Minnesota Press, 1970.
Hameroff, S. R. "Quantum coherence in microtubules: A neural basis for an emergent consciousness?", *Journal of Consciousness Studies, 1*, 1994, 91-118.
Hardin, C. L. "Qualia and materialism: Closing the explanatory gap", *Philosophy and Phenomenological Research, 48*, 1987, 281-98.
Hardin, C. L. *Color for Philosophers: Unweaving the Rainbow*, Indianapolis, Hackett, 1988.

Hare, R. M. *The Language of Morals*, Oxford, Clarendon Press, 1952.
Hare, R. M. "Supevenience", *Proceedings of the Aristotelian Society, supl., 58*, 1984, 1-16.
Harman, G. "Conceptual role semantics", *Notre Dame Journal of Formal Logic, 28,* 1982, 242-56.
Harman, G. "The intrinsic quality of experience", *Philosophical Perspectives, 4*, 1990, 31-52.
Harnad, S. "Minds, machines and Searle", *Journal of Experimental and Theoretical Artificial Intelligence, 1*, 1989, 5-25.
Harrison, B. "On describing colors", *Inquiry, 10*, 1967, 38-52.
Harrison, B. *Form and Content*, Oxford, Blackwell, 1973.
Harrison, J. "Three philosophical fairy stories", *Ratio, 23*, 1981a, 63-67.
Harrison, J. "Gulliver's adventures in Fairyland", *Ratio, 23*, 1981b, 158-64.
Haugeland, J. "Programs, causal powers, and intentionality", *Behavioral and Brain Sciences, 4*, 1980, 432-33.
Haugeland, J. "Weak supervenience", *American Philosophical Quarterly, 19,* 1982, 93-103.
Healey, R. A. "How many worlds?", *Nous, 18*, 1984, 591-616.
Heil, J. *The Nature of True Minds*, Cambridge, Cambridge University Press, 1992.
Hellman, G. y F. Thompson. "Physicalism: Ontology, determination and reduction, *Journal of Philosophy, 72*, 1975, 551-64.
Hill, C. S. *Sensations: A Defense of Type Materialism*, Cambridge, Cambridge University Press, 1991.
Hodgson, D. *The Mind Matters: Consciousness and Choice in a Quantum World*, Oxford, Oxford University Press, 1988.
Hofstadter, D. R. *Gödel, Escher, Bach: an Eternal Golden Braid*, Nueva York, Basic Books, 1979. [*Gödel, Escher, Bach.* Barcelona, Tusquets, 3ª ed., 1989.]
Hofstadter, D. R. "Reflections on Searle", en D.R. Hofstadter y D.C. Dennett (comps.), *The Mind's I*, Nueva York, Basic Books, 1981.
Hofstadter, D. R. "Who shoves whom around inside the careenium?, en *Metamagical Themas*, Nueva York, Basic Books, 1985a.
Hofstadter, D. R. "Heisenberg's Uncertainty Principle and the many-worlds interpretation of quantum mechanics, en *Metamagical Themas*, Nueva York, Basic Books, 1985b.
Honderich, T. "Psychophysical law-like connections and their problems, *Inquiry*, *24*, 1981, 277-303.
Horgan, T. "Supervenient bridge laws", *Philosophy of Science, 45,* 1978, 227-49.
Horgan, T. "Supervenience and microphysics", *Pacific Philosophical Quarterly, 63*, 1982, 29-43.
Horgan, T. "Functionalism, qualia, and the inverted spectrum", *Philosophy and Phenomenological Research*, *44*, 1984a, 453-69.
Horgan, T. "Jackson on physical information and qualia", *Philosophical Quarterly, 34*, 1984b, 147-83.
Horgan, T. "Supervenience and cosmic hermeneutics", *Southern Journal of Philosophy,* supl. *22*, 1984c, 19-38.

Horgan, T. "Supervenient qualia", *Philosophical Review, 96*, 1987, 491-520.
Horgan, T. "From supervenience to superdupervenience: Meeting the demands of a material world", *Mind, 102*, 1993, 555-86.
Horgan, T. y M. Timmons. "Troubles for new wave moral semantics: The 'open question argument' revived", *Philosophical Papers*, 1992a.
Horgan, T. y M. Timmons. "Trouble on moral twin earth: Moral queerness revived", *Synthese, 92,* 1992b, 223-60.
Horst, S. "Phenomenology and psychophysics", manuscrito, Wesleyan University, 1995.
Hughes, R. I. G. *The Structure and Interpretation of Quantum Mechanics*, Cambridge, Mass., Harvard University Press, 1989.
Humphrey, N. *A History of the Mind: Evolution and the Birth of Consciousness*, Nueva York, Simon and Schuster, 1992. [*Una historia de la mente: la evolución y el nacimiento de la conciencia.* Barcelona, Gedisa, 1995.]
Huxley, T. "On the hypothesis that animals are automata", en *Collected Essays*, Londres, 1874, 1893-94.
Jackendoff, R. *Consciousness and the Computational Mind*, Cambridge, Mass., MIT Press, 1987.
Jackson, F. *Perception*, Cambridge, Cambridge University Press, 1977.
Jackson, F. "A note on physicalism and heat", *Australasian Journal of Philosophy, 58*, 1980, 26-34.
Jackson, F. "Epiphenomenal qualia", *Philosophical Quarterly, 32,* 1982, 127-36.
Jackson, F. "Armchair metaphysics", en J. O'Leary-Hawthorne y M. Michael (comps.), *Philosophy in Mind*, Dordrecht, Kluwer, 1993.
Jackson, F. "Finding the mind in the natural world", en R. Casati, B. Smith y G. White (comps.), *Philosphy and the Cognitive Sciences*, Viena, Holder-Pichler-Tempsky, 1994.
Jackson, F. "Postscript to 'What Mary didn't know'", en P.K. Moser y J.D. Trout (comps.), *Contemporary Materialism*, Londres, Routledge, 1995.
Jacoby, H. "Empirical functionalism and conceivability arguments", *Philosophical Psychology*, *2*, 1990, 271-82.
Jaynes, J. *The Origins of Consciousness in the Breakdown of the Bicameral Mind*, Boston, Houghton Mifflin, 1976.
Johnson-Laird, P. "A computational analysis of consciousness", *Cognition and Brain Theory*, *6,* 1983, 499-508.
Kaplan, D. *Dthat*, en P. Cole (comp.), *Syntax and Semantics*, Nueva York, Academic Press, 1979.
Kaplan, D. "Demonstratives", en J. Almog, J. Perry y H. Wettstein (comps.), *Themes from Kaplan*, Nueva York, Oxford University Press, 1989.
Kim, J. "Supervenience and nomological incommensurables", *American Philosophical Quarterly, 15*, 1978, 149-56.
Kim, J. "Concepts of supervenience", *Philosophy and Phenomenological Research, 45*, 1984, 153-76.
Kim, J. "Psychophysical laws", en B. McLaughlin y E. LePore (comps.), *Action and Events*, Oxford, Blackwell, 1985.
Kim, J. "Mechanism, purpose, and explanatory exclusion", *Philosophical Perspectives*, *3*, 1989, 77-108.

Kim, J. *Supervenience and Mind*, Cambridge, Cambridge University Press, 1993.
Kirk, R. "Zombies versus materialists", *Aristotelian Society, 48* (supl.), 1974, 135-52.
Kirk, R. "From physical explicability to full-blooded materialism", *Philosophical Quarterly, 29*, 1979, 229-37.
Kirk, R. "Consciousness and concepts", *Proceedings of the Aristotelian Society, 66* (supl.), 1992, 23-40.
Kirk, R. *Raw Feeling: A Philosophical Account of the Essence of Consciousness*, Oxford, Oxford University Press, 1994.
Korb, K. "Searle's AI program", *Journal of Experimental and Theoretical Artificial Intelligence, 3*, 1991, 283-96.
Kripke, S.A. "Identity and necessity", en M. Munitz (comp.), *Identity and Individuation*, Nueva York, New York University Press, 1971.
Kripke, S.A. "Naming and necessity", en G. Harman y D. Davidson (comps.), *The Semantics of Natural Language*, Dordrecht, Reidel, 1972. [Reimpreso como Kripke, 1980.]
Kripke, S.A. *Naming and Necessity*, Cambridge, Mass., Harvard University Press, 1980.
Kripke, S.A. *Wittgenstein on Rule-Following and Private Language*, Cambridge, Mass., Harvard University Press, 1982.
Lahav, R. y N. Shanks. "How to be a scientifically respectable 'property dualist'", *Journal of Mind and Behavior, 13*, 1992, 211-32.
Langton, C. G. *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems,* Redwood City, Calif., Addison-Wesley, 1989.
Leckey, M. "The universe as a computer: A model for prespace metaphysics", manuscrito, Philosophy Department, Monash University, 1993.
Levine, J. "Materialism and qualia: The explanatory gap", *Pacific Philosophical Quarterly, 64*, 1983, 354-61.
Levine, J. "Absent and inverted qualia revisited", *Mind and Language, 3,* 1988, 271-87.
Levine, J. "Cool red", *Philosophical Psychology, 4,* 1991, 27-40.
Levine, J. "On leaving out what it's like", en M. Davies y G. Humphreys (comps.), *Consciousness: Psychological and Philosophical Essays,* Oxford, Blackwell, 1993.
Lewis, D. "An argument for the identity theory", *Journal of Philosophy, 63*, 1966, 17-25.
Lewis, D. "Psychophysical and theoretical identifications", *Australasian Journal of Philosophy, 50*, 1972, 249-58.
Lewis, D. *Counterfactuals*, Cambridge, Mass., Harvard University Press, 1973.
Lewis, D. "Radical interpretation", *Synthese, 23*, 1974, 331-44.
Lewis, D. "Attitudes *de dicto* and *de re*", *Philosophical Review, 88*, 1979, 513-45.
Lewis, D. "Extrinsic properties", *Philosophical Studies, 44,* 1983a, 197-200.
Lewis, D. "New work for a theory of universals", *Australasian Journal of Philosophy, 61*, 1983b, 343-77.

Lewis, D. *On the Plurality of Worlds*, Oxford, Blackwell, 1986a.
Lewis, D. *Philosophical Papers*, Vol. 2, Nueva York, Oxford University Press, 1986b.
Lewis, D. "What experience teaches", en W. Lycan (comp.), *Mind and Cognition*, Oxford, Blackwell, 1990.
Lewis, D. "Reduction of mind", en S. Guttenplan (comp.), *A Companion to the Philosophy of Mind*, Oxford, Blackwell, 1994.
Libet, B. "The neural time factor in conscious and unconscious events", en *Experimental and Theoretical Studies of Consciousness*, Ciba Foundation Symposium 174, Nueva York, Wiley, 1993.
Loar, B. "Phenomenal states", *Philosophical Perspectives, 4*, 1990, 81-108.
Lockwood, M. *Mind, Brain, and the Quantum*, Oxford, Blackwell, 1989.
Lockwood, M. "The grain problem", en H. Robinson (comp.), *Objections to Physicalism*, Oxford, Oxford University Press, 1992.
Logothetis, N. y J. D. Schall. "Neuronal correlates of subjective visual perception", *Science, 245*, 1989, 761-763.
London, F. y E. Bauer. "The theory of observation in quantum mechanics" (en francés), *Actualités scientifiques et industrielles,* No. 775, 1939. [Traducción inglesa en Wheeler y Zurek, 1983.]
Lucas, J. R. "Minds, machines and Gödel", *Philosophy, 36*, 1961, 112-27.
Lycan, W. G. "Inverted spectrum", *Ratio, 15*, 1973, 315-319.
Lycan, W. G. *Consciousness*, Cambridge, Mass., MIT Press, 1987.
Lycan, W. G. "A limited defense of phenomenal information", en T. Metzinger (comp.), *Conscious Experience*, Paderborn, Schöningh, 1995.
Lycan, W. G. *Consciousness and Experience*, Cambridge, Mass., MIT Press, 1996.
Mackay, D. M. *Information, Mechanism, and Meaning*, Cambridge, Mass., MIT Press, 1969.
Mackie, J. L. *The Cement of the Universe,* Oxford, Oxford University Press, 1974.
Mackie, J. L. *Ethics: Inventing Right and Wrong*, Harmondsworth, Penguin Books, 1977.
Marks, L. E. *The Unity of the Senses: Interrelations among the Modalities*, Nueva York, Academic Press, 1978.
Matzke, D. (comp.). *Proceedings of the 1992 Workshop on Physics and Computation*, Los Alamitos, California, IEEE Computer Society Press, 1993.
Matzke, D. (comp.) *Proceedings of the 1994 Workshop on Physics and Computation*, Los Alamitos, California, IEEE Computer Society Press, 1995.
Maxwell, G. "Rigid designators and mind-brain identity", en C.W. Savage (comp.), *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science, Vol. 9, Minneapolis, University of Minnesota Press, 1978.
McCarthy, J. "Ascribing mental qualities to machines", en M. Ringle (comp.), *Philosophical Perspectives in Artificial Intelligence*, Atlantic Highlands, N.J., Humanities Press, 1979.
McDowell, J. *Mind and World*, Cambridge, Mass., Harvard University Press, 1994.

McGinn, C. "Anomalous monism and Kripke's Cartesian intuitions", *Analysis, 2*, 1977, 78-80.
McGinn, C. "Can we solve the mind-body problem?", *Mind, 98*, 1989, 349-66.
McLaughlin, B. P. "The rise and fall of the British emergentists", en A. Beckermann, H. Flohr y J. Kim (comps.), *Emergence or Reduction? Prospects for Nonreductive Physicalism*, Berlín, De Gruyter, 1992.
McLaughlin, B. P. "Varieties of supervenience", en E. E. Savellos y U. D. Yalcin (comps.), *Supervenience: New Essays*, Cambridge, Cambridge University Press, 1995.
McMullen, C. "'Knowing what it's like' and the essential indexical", *Philosophical Studies, 48*, 1985, 211-33.
Meehl, P. E. y W. Sellars. "The concept of emergence", en H. Feigl y M. Scriven (comps.), *The Foundations of Science and the Concept of Psychology and Psychoanalysis*, Minnesota Studies in the Philosophy of Science, vol. 1, Minneapolis, University of Minnesota Press, 1956.
Molnar, G. "Kneale's argument revisited", *Philosophical Review, 78*, 1969, 79-89.
Moore, G. E. *Philosophical Studies*, Londres, Routledge and Kegan Paul, 1922.
Müller, G. E. "Zur Psychophysik der Gesichtsempfindungen", *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, 10, 1896, 1-82.
Nagel, T. "Armstrong on the mind", *Philosophical Review, 79*, 1970, 394-403.
Nagel, T. "What is it like to be a bat?", *Philosophical Review, 4*, 1974, 435-50.
Nagel, T. "The objective self", en C. Ginet y S. Shoemaker (comps.), *Knowledge and Mind: Philosophical Essays,* Nueva York, Oxford University Press, 1983.
Nagel, T. *The View from Nowhere*, Nueva York, Oxford University Press, 1986.
Natsoulas, T. "Consciousnes", *American Psychologist, 33*, 1978, 906-14.
Nelkin, N. "Unconscious sensations", *Philosophical Psychology, 2*, 1989, 129-41.
Nelkin, N. "What is consciousness?", *Philosophy of Science, 60*, 1993, 419-34.
Nemirow, L. "Physicalism and the cognitive role of acquaintance", en W. Lycan (comp.), *Mind and Cognition*, Oxford, Blackwell, 1990.
Newell, A. "SOAR as a unified theory of cognition: Issues and explanations", *Behavioral and Brain Sciences, 15,* 1992, 464-92.
Newton, N. "Machine understanding and the Chinese Room", *Philosophical Psychology, 2*, 1989, 207-15.
Nida-Rümelin, M. "What Mary couldn't know: Belief about phenomenal states", en T. Metzinger (comp.), *Conscious Experience*, Paderborn, Schöningh, 1995.
Nida-Rümelin, M. "Pseudonormal vision: An actual case of qualia inversion?", *Philosophical Studies,* 1996.
Papineau, D. *Philosophical Naturalism*, Oxford, Blackwell, 1993.
Parfit, D. *Reasons and Persons,* Oxford, Oxford University Press, 1984.
Peacocke, C. "Scenarios, concepts, and perception", en T. Crane (comp.). *The Contents of Experience*, Cambridge, Cambridge University Press, 1992.
Penrose, R. "Quantum physics and conscious thought", en B. Hiley y Peat

(comps.), *Quantum Implications: Essays in Honor of David Bohm*, Nueva York, Methuen, 1987.

Penrose, R. *The Emperor's New Mind*, Oxford, Oxford University Press, 1989. [*La nueva mente del emperador*. Madrid, Círculo de Lectores, 1996; *La nueva mente del emperador*. Barcelona, Grijalbo, 1996; *La nueva mente del emperador*. Barcelona, Mondadori España, 1991.]

Penrose, R. *Shadows of the Mind*, Oxford, Oxford University Press, 1994. [*Las sombras de la mente*. Barcelona, Crítica, 1996.]

Perry, J. "The problem of the essential indexical", *Nous, 13,* 1979, 3-21.

Petrie, B. "Global supervenience and reduction", *Philosophy and Phenomenological Research, 48*, 1987, 119-30.

Place, U. T. "Is consciousness a brain process?", *British Journal of Psychology, 47*, 1956, 44-50.

Plantinga, A. "Actualism and possible worlds", *Theoria, 42,* 1976, 139-60.

Putnam, H. "Minds and machines", en S. Hook (comp.), *Dimensions of Mind,* Nueva York, New York University Press, 1960.

Putnam, H. "The meaning of 'meaning'", en K. Gunderson (comp.), *Language, Mind, and Knowledge*, Minneapolis, University of Minnesota Press, 1975.

Putnam, H. *Reason, Truth, and History*, Cambridge, Cambridge University Press, 1981. [*Razón, verdad e historia*. Madrid, Tecnos, 1988.]

Putnam, H. "Possibility and necessity", en *Philosophical Papers*, vol. 3, Cambridge, Cambridge University Press, 1983.

Putnam, H. *Representation and Reality*, Cambridge, Mass., MIT Press, 1988. [*Representación y realidad: Un balance crítico del funcionalismo*. Barcelona, Gedisa, 1990.]

Pylyshyn, Z. "The 'causal power' of machines", *Behavioral and Brain Sciences, 3*, 1980, 442-44.

Quine, W. V. "Two dogmas of empiricism", *Philosophical Review, 60*, 1951, 20-43.

Quine, W. V. "Propositional objects", en *Ontological Relativity and Other Essays*, Nueva York, Columbia University Press, 1969. [*La relatividad ontológica y otros ensayos*. Madrid, Tecnos, 1974.]

Rensink, R. A., J. K. O'Reagan y J. J. Clark. "Image flicker is as good as saccades in making large scene changes invisible", *Perception, 24* (supl.), 1995, 26-27.

Rey, G. "A reason for doubting the existence of consciousness", en R. Davidson, S. Schwartz y D. Shapiro (comps.), *Consciousness and Self-Regulation*, Vol. 3, Nueva York, Plenum.

Rey, G. "What's really going on in Searle's 'Chinese Room'", *Philosophical Studies, 50*, 1986, 169-85.

Rey, G. "Sensational sentences reversed", *Philosophical Studies, 68*, 1992, 289-319.

Reynolds, C. "Flocks, herds, and schools: A distributed behavioral model", *Computer Graphics, 21,* 1987, 25-34.

Robinson, H. "The mind-body problem in contemporary philosophy", *Zygon, 11*, 1976, 346-60.

Robinson, H. *Matter and Sense*, Cambridge, Cambridge University Press, 1982.

Robinson, W. *Brains and People: An Essay on Mentality and Its Causal conditions*, Filadelfia, Temple University Press, 1988.
Rosenberg, G. H. "Consciousness and causation: Clues toward a double-aspect theory", manuscrito, Indiana University, 1996.
Rosenthal, D. M. "A theory of consciousness", en N. Block, O. Flanagan y G. Güzeldere (comps.), *The Nature of Consciousness*, Cambridge, Mass., MIT Press, 1996.
Russell, B. *The Analysis of Matter,* Londres, Kegan Paul, 1927. [*Análisis de la materia*, Barcelona, Taurus, 2ª ed., 1976.]
Ryle, G. *The concept of Mind*, Londres, Hutchinson, 1949.
Savage, C.W. *The Measurement of Sensation*, Berkeley, University of California Press, 1970.
Savitt, S. "Searle's demon and the brain simulator reply", *Behavioral and Brain Sciences, 5*, 1982, 342-43.
Sayre, K. M. *Cybernetics and the Philosophy of Mind*, Atlantic Highlands, N.J., Humanities Press, 1976.
Sayre-McCord, G. "Introduction: The many moral realisms", en G. Sayre-McCord (comp.), *Essays on Moral Realism*, Ithaca, N.Y., Cornell University Press, 1988.
Schacter, D. L. "On the relation between memory and consciousness: Dissociable interactions and conscious experience", en H. Roediger y F. Craik (comps.), *Varieties of Memory and consciousness: Essays in Honor of Endel Tulving*, Hillsdale, N.J., Erlbaum, 1989.
Schlick, M. "Positivism and Realism", *Erkenntnis, 3,* 1932.
Schlick, M. "Form and content: An introduction to philosophical thinking", en *Gesamelte Aufsätze 1926-1936*, Viena, Gerold, 1938. [Reimpreso en H. L. Mulder y B. van de Velde-Schlick (comps.), *Philosophical Papers*, Dordrecht, Reidel, 1979.]
Seager, W. E. "Weak supervenience and materialism", *Philosophy and Phenomenological Research*, *48*, 1988, 697-709.
Seager, W. E. *Metaphysics of Consciousness*, Londres, Routledge, 1991.
Searle, J. R. "Minds, brains, and programas", *Behavioral and Brain Sciences, 3*, 1980, 417-24.
Searle, J. R. *Minds, Brains and Science*, Cambridge, Mass., Harvard University Press, 1984. [*Mentes, cerebros y ciencia.* Madrid, Cátedra, 2ª ed., 1985.]
Searle, J. R. "Consciousness, explanatory inversion and cognitive science", *Behavioral and Brain Sciences, 13*, 1990a, 585-642.
Searle, J. R. "Is the brain a digital computer?", *Proceedings and Addresses of the American Philosophical Association,* 64, 1990b, 21-37.
Searle, J. R. *The Rediscovery of the Mind*, Cambridge, Mass., MIT Press, 1992. [*El descubrimiento de la mente.* Barcelona, Crítica, 1996.]
Sellars, W. "Empiricism and the philosophy of mind", en H. Feigl y M. Scriven (comps.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, Minnesota Studies in the Philosophy of Science, vol. 1, Minneapolis, University of Minnesota Press, 1956.
Sellars, W. "The identity approach to the mind-body problem", *Review of Metaphysics, 18*, 1965, 430-51.
Sellars, W. "Is consciousness physical?", *Monist, 64*, 1981, 66-90.

Shallice, T. "Dual functions of consciousness", *Psychological Review*, 79, 1972, 383-93.
Shallice, T. "Information-processing models of consciousness: Possibilities and problems", en A. Marcel y E. Bisiach (comps.), *Consciousness in Contemporary Science*, Oxford, Oxford University Press, 1988a.
Shallice, T. *From Neuropsychology to Mental Structure*, Cambridge, Cambridge University Press, 1988b.
Shannon, C.E. "A mathematical theory of communication", *Bell Systems Technical Journal, 27*, 1948, 379-423. [Reimpreso en C.E. Shannon y W. Weaver. *The Mathematical Theory of Communication,* Urbana, University of Illinois Press, 1949.] [*Teoría matemática de la comunicación.* Madrid, Forja, 1981.]
Shepard, R. N. "On the physical basis, linguistic representation, and conscious experience of colors", en G. Harman (comp.), *Conceptions of the Human Mind: Essays in Honor of George A. Miller*, Hillsdale, N.J., Erlbaum, 1993.
Shoemaker, S. "Functionalism and qualia", *Philosophical Studies, 27*, 1975a, 291-315.
Shoemaker, S. "Phenomenal similarity", *Critica, 7*, 1975b, 3-37.
Shoemaker, S. "Causality and properties", en P. van Inwagen (comp.), *Time and Cause*, Dordrecht, Reidel, 1980.
Shoemaker, S. "Some varieties of functionalism", *Philosophical Topics, 12*, 1981, 93-119.
Shoemaker, S. "The inverted spectrum", *Journal of Philosophy, 79*, 1982, 357-81.
Sidelle, A. *Necessity, essence, and individuation*, Ithaca, N.Y., Cornell University Press, 1989.
Sidelle, A. "Rigidity, ontology, and semantic structure", *Journal of Philosophy, 8,* 1992, 410-30.
Siewert, C. "What Dennett can't imagine and why", *Inquiry, 36*, 1993, 93-112.
Siewert, C. "Understanding consciousness", Tesis de doctorado, University of California, Berkeley. [Próximamente como libro de Princeton University Press.]
Skyrms, B. *Causal Necessity*, New Haven, Yale University Press, 1980.
Smart, J. J. C. "Sensations and brain processes", *Philosophical Review, 68,* 1959, 141-56.
Sperling, G. "The information available in brief visual presentations", *Psychological Monographs, 74*, 1960.
Sperry, R. W. "A modified concept of consciousness", *Psychological Review, 76*, 1969, 532-36.
Sperry, R. W. "Turnabout on consciousness: A mentalist view, *Journal of Mind and Behavior, 13*, 1992, 259-80.
Sprigge, T. L. S. "Final causes", *Proceedings of the Aristotelian Society*, 45 (supl.), 1971, 149-70.
Sprigge, T. L. S. "Consciousness", *Synthese, 98*, 1994, 73-93.
Squires, E. *Conscious Mind in the Physical World*, Bristol, Hilger, 1990.
Stalnaker, R. "Possible worlds", *Nous, 10,* 1976, 65-75.

Stalnaker, R. "Assertion", en P. Cole (comp.), *Syntax and Semantics: Pragmatics,* vol. 9, Nueva York, Academic Press, 1978.
Stapp, H. P. *Mind, Matter, and Quantum Mechanics*, Berlín, Springer-Verlag, 1993.
Stevens, S. S. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*, Nueva York, Wiley, 1975.
Sutherland, N. S. (comp.) *The International Dictionary of Psychology,* Nueva York, Continuum, 1989.
Swoyer, C. "The nature of natural laws", *Australasian Journal of Philosophy, 60*, 1982, 203-23.
Teller, D. Y. "Linking propositions", *Vision Research, 24,* 1984, 1233-46.
Teller, D. Y. "The domain of visual science", en L. Spillman y J.S. Werner (comps.), *Visual Perception: The Neurophysiological Foundations*, Nueva York, Academic Press, 1990.
Teller, P. "A poor man's guide to supervenience and determination", *Southern Journal of Philosophy*, supl., *22,* 1984, 137-62.
Teller, P. "A contemporary look at emergence", en A. Beckermann, H. Flohr y J. Kim (comps.), *Emergence or Reduction? Prospects for Nonreductive Physicalism,* Berlín, De Gruyter, 1992.
Thagard, P. "The emergence of meaning: An escape from Searle's Chinese Room", *Behaviorism, 14*, 1986, 139-46.
Thompson, E. "Novel colours", *Philosophical Studies, 68*, 1992, 321-49.
Tienson, J. L. "Brains are not conscious", *Philosophical Papers, 16*, 1987, 187-93.
Tooley, M. "The nature of laws", *Canadian Journal of Philosophy*, 7, 1977, 667-98.
Tooley, M. *Causation: A Realist Approach*, Oxford, Oxford University Press, 1987.
Tye, M. "The subjective qualities of experience", *Mind, 95*, 1986, 1-17.
Tye, M. "Visual qualia and visual content", en T. Crane (comp.), *The Contents of Experience*, Cambridge, Cambridge University Press, 1992.
Tye, M. "Blindsight, the absent qualia hypothesis, and the mystery of consciousness", en C. Hookway (comp.), *Philosophy and the Cognitive Sciences*, Cambridge, Cambridge University Press, 1993.
Tye, M. *Ten Problemas of Consciousness*, Cambridge, Mass., MIT Press, 1995.
Van Cleve, J. "Mind-dust or magic? Panpsychism versus emergence", *Philosophical Perspectives, 4*, 1990, 215-26.
Van Gulick, R. "A funcionalist plea for self-consciousness", *Philosophical Review, 97*, 1988, 149-88.
Van Gulick, R. "Nonreductive materialism and the nature of intertheoretical constraint", en A. Beckermann, H. Flohr y J. Kim (comps.), *Emergence or Reduction? Prospects for Nonreductive Physicalism*, Berlín, De Gruyter, 1992.
Van Gulick, R. "Understanding the phenomenal mind: Are we all just armadillos?", en M. Davies y G. Humphreys (comps.), *Consciousness: A Mind and Language Reader*, Oxford, Blackwell, 1993.
Velmans, M. "Is human information processing conscious?", *Behavioral and Brain Sciences, 14*, 1991, 651-69.

Weinberg, S. *Dreams of a Final Theory*, Nueva York, Pantheon Books, 1992. [*El sueño de una teoría final*. Barcelona, Crítica, 1994.]
Weiskrantz, L. *Blindsight: A Case Study and Implications*, Oxford, Oxford University Press, 1986.
Weiskrantz, L. "Introduction: Dissociated issues", en D. Milner y M. Rugg (comps.), *The Neuropsychology of Consciousness*, Londres, Academic Press.
Wheeler, J. A. "Information, physics, quantum: The search for links", en W.H. Zurek (comp.), *Complexity, Entropy, and the Physics of Information*, Redwood City, Calif., Addison-Wesley, 1990.
Wheeler, J. A. "It from bit", en *At Home in the Universe*, Woodbury, N.Y., American Institute of Physics Press, 1994.
Wheeler, J. A. y W. H. Zurek. *Quantum Theory and Measurement*, Princeton, Princeton University Press, 1983.
White, S. L. "Curse of the qualia", *Synthese, 68*, 1986, 333-68.
Wigner, E. P. "Remarks on the mind-body question", en I.J. Good (comp.), *The Scientist Speculates*, Nueva York, Basic Books.
Wilkes, K. V. "Is consciousness important?", *British Journal for the Philosophy of Science*, *35*, 1984, 223-43.
Wilson, M. "Predicate meets property", *Philosophical Review, 91*, 1982, 549-89.
Wilson, M. What is this thing called "pain"? The philosophy of science behind the contemporary debate", *Pacific Philosophical Quarterly, 66*, 1985, 227-67.
Winograd, T. *Understanding Natural Language,* Nueva York, Academic Press, 1972.
Wittgenstein, L. *Philosophical Investigations*, Londres, Macmillan, 1953. [*Investigaciones filosóficas*. Barcelona, Crítica, 1988.]
Wittgenstein, L. "Notes for lectures on 'private experience' and 'sense data'", *Philosophical Review, 77*, 1965.
Wright, R. *Three Scientists and Their Gods*, Nueva York, Times Books, 1988.
Yablo, S. "Is conceivability a guide to possibility?", *Philosophy and Phenomenological Research, 53*, 1993, 1-42.
Zuboff, A. "What is a mind?", en P. A. French, T. E. Uehling y H. K. Wettstein (comps.), *Philosophical Naturalism*, Midwest Studies in Philosophy, vol. 19, Notre Dame, Ind., University of Notre Dame Press, 1994.
Zurek, W. H. *Complexity, Entropy, and the Physics of Information*, Redwood City, Calif., Addison-Wesley, 1990.

# Índice temático



Ciencias cognitivas
Serie CLA·DE·MA

gedisa
editorial

# David J. Chalmers
# La mente consciente

¿Qué es la conciencia? ¿Cómo pueden los procesos físicos en el cerebro dar lugar a la actividad de una mente consciente? Estas cuestiones suscitan las discusiones más acaloradas en el ámbito de la filosofía y la ciencia actuales. El filósofo David J. Chalmers ofrece aquí un análisis conciso de este debate y esboza una nueva teoría sobre la conciencia, que rechaza las tendencias reduccionistas y que sigue siendo compatible con las concepciones fisicalistas, pero va más allá de ellas.

El autor nos acompaña en un viaje de gran alcance por las ramificaciones de las concepciones filosóficas sobre la conciencia. Al contrario de las ciencias cognitivas y la neurociencia contemporáneas, Chalmers propone entender la experiencia consciente como una entidad irreductible -al igual que las propiedades físicas de tiempo, masa y espacio-, que se encontraría en niveles muy profundos y que no se puede entender como la mera suma de componentes físicos más simples.

En la segunda mitad del libro, el autor construye una teoría fundamental acerca de las leyes básicas que gobiernan la estructura y el carácter de las experiencias conscientes, mostrando cómo esta reconceptualización de la mente podría llevarnos a una nueva ciencia de la conciencia.

«Tomando como punto de partida unas nociones muy intuitivas sobre la conciencia, David Chalmers llega a conclusiones sorprendentes sobre lo que realmente es el núcleo central de la existencia humana. Se trata de una importante exploración del tema, brillantemente argumentada desde un conocimiento perfecto del territorio. Aunque personalmente no puedo acompañar a Chalmers siempre a donde me quiere llevar, es ciertamente uno de los mejores guías posibles.»

*Douglas Hofstadter*, Indiana University

**David J. Chalmers** es profesor de filosofía de la Universidad de Santa Cruz, California. Nació en Sydney y trabajó como *Rhodes Scholar* en la Universidad de Oxford y como *McDonnell Fellow* en la Universidad de Washington. Su artículo "The puzzle of Conscious Experience" apareció como número extraordinario del *Scientific American*.

ISBN 84-7432-692-3



9 788474 326925
302475